ERGO Bioinformatics Suite:
tools for microbial genome analysis

BKD Meeting, November 2005

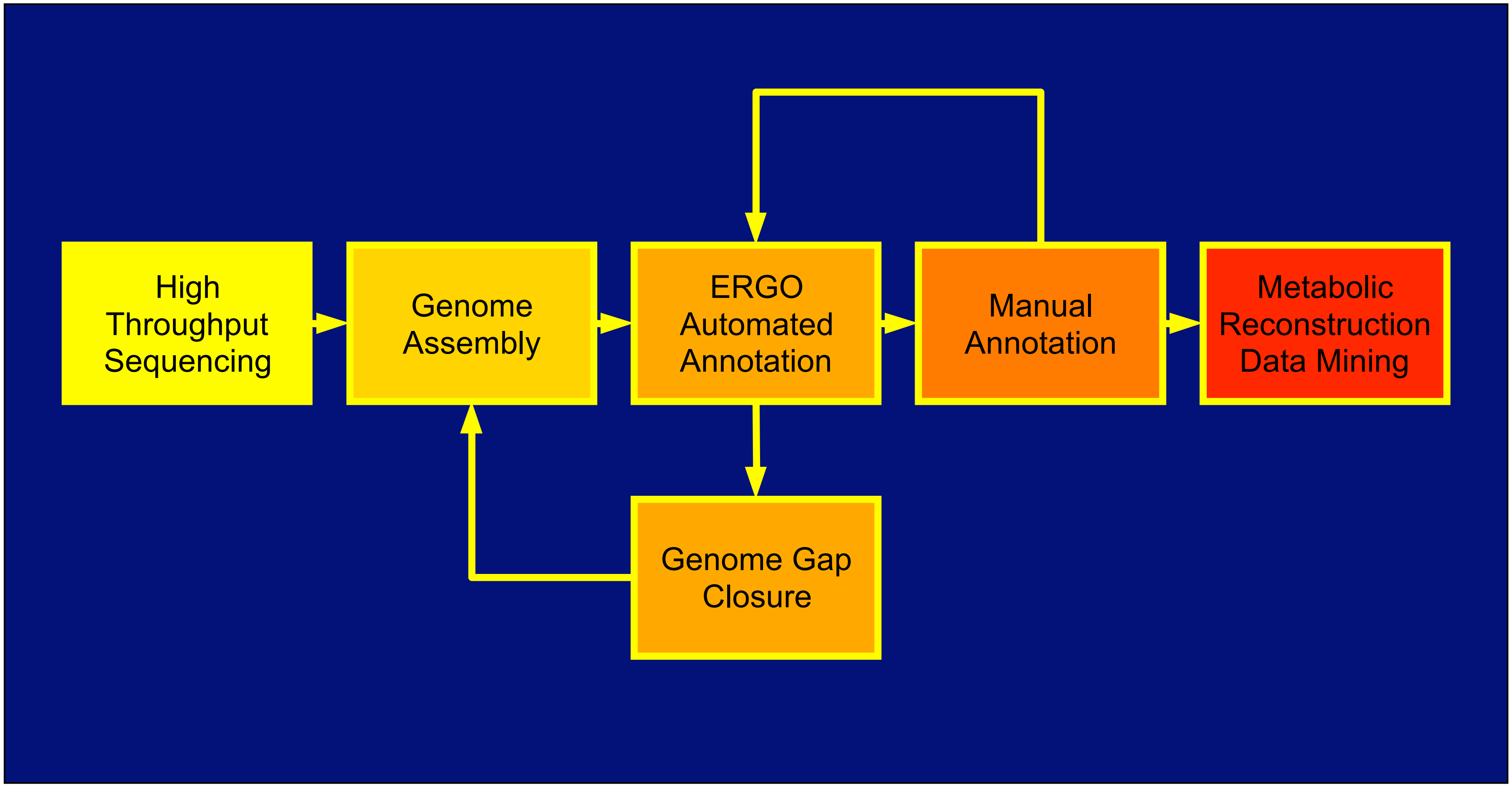
Anamitra Bhattacharyya, Ph.D.
Integrated Genomics, Inc.
anamitra@integratedgenomics.com

Agenda

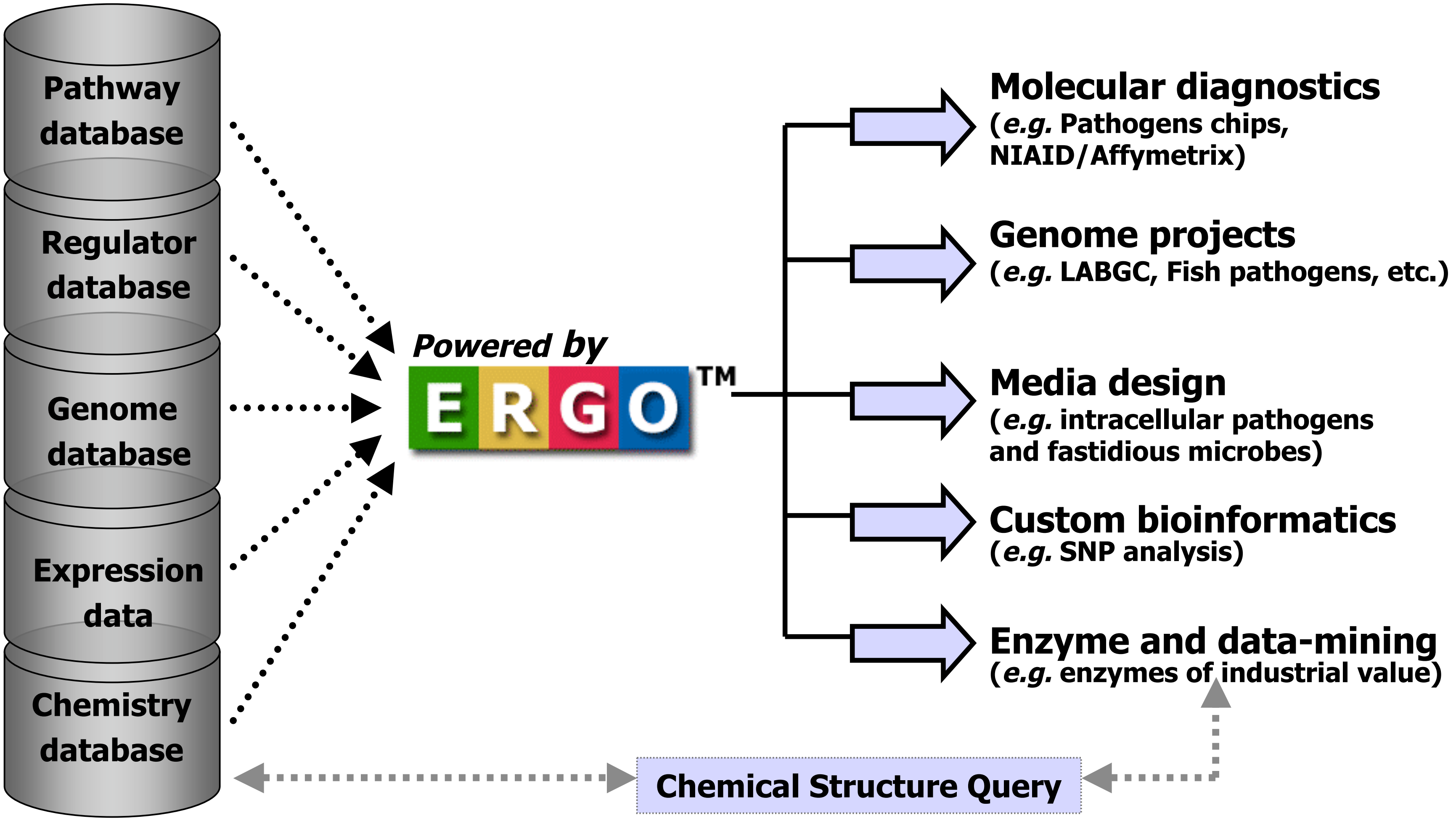
ERGO^{TM} Genome Analysis Platform

- **ERGO Contents, Utilities and Applications**
- **Application Case Study 1: Genome comparison of multiple *Xylella* strains**
- **Application Case Study 2: Genetic basis for phenotypic traits (*Fusobacteria*)**
- **Initial Observations: *R. salmoninarum* genome**

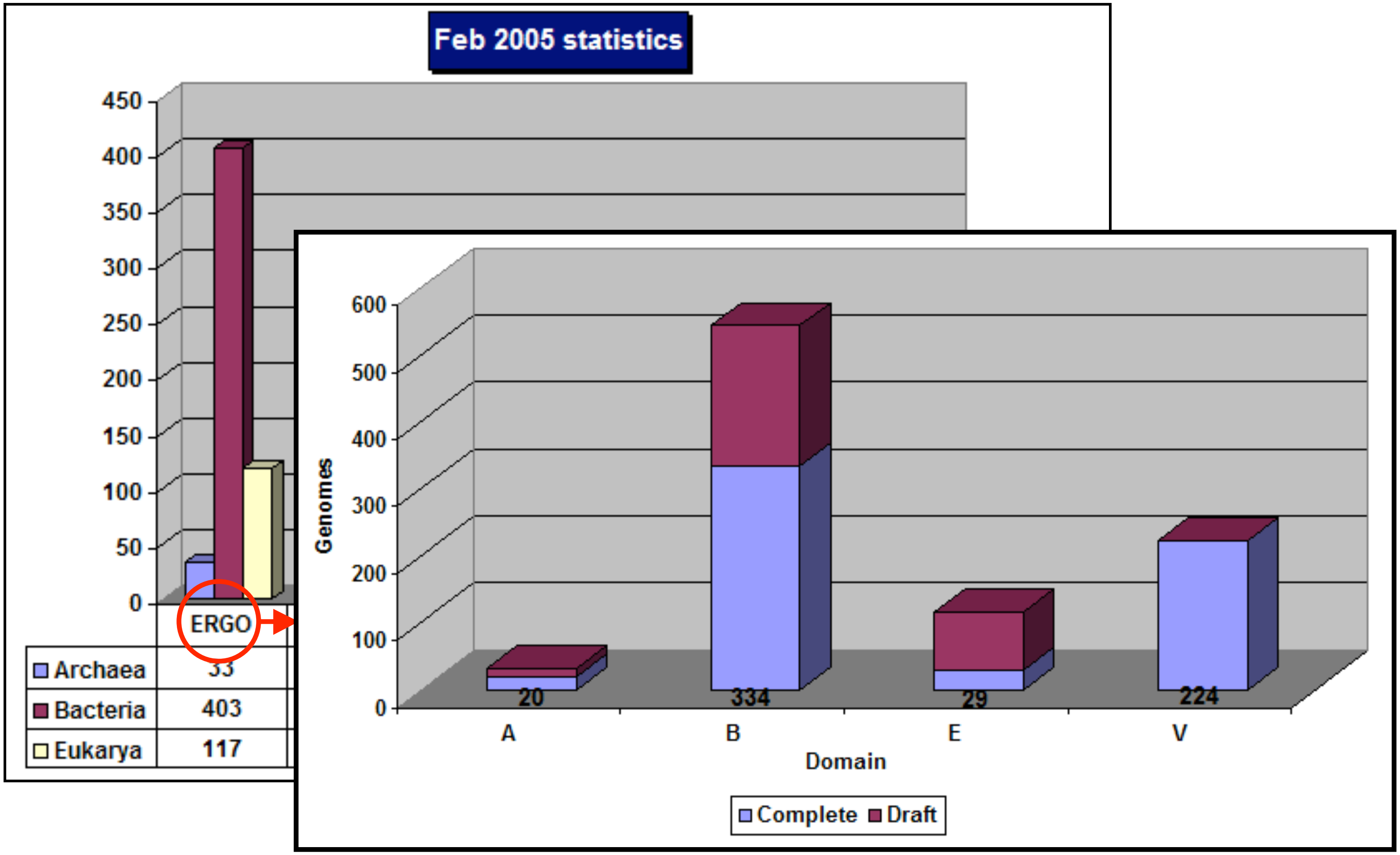
ERGO is Central to Getting the Most Out of a Genome Project


High Throughput Sequencing
Genome Assembly
ERGO Automated Annotation
Manual Annotation
Metabolic Reconstruction Data Mining
Genome Gap Closure

ERGO™ Genome Analysis and Discovery Suite: biological platform for *in silico* analysis



ERGO: largest and diverse genome inventory



	ERGO
▨ Archaea	33
▨ Bacteria	403
□ Eukarya	117

ERGO: A Unique Repository of Genomic Data

- **Protein-encoding genes:**
  - **SwissProt**: **133,000** carefully curated sequences
  - **PIR:** **1.1 million** less curated sequences
  - **GenBank**: **1.5 million** far less curated sequences
  - ERGO: 2.2 million carefully curated sequences (in non-redundant protein sequence database)

ERGO curated Protein-encoding genes: 1.4 million sequences

ERGO Statistics

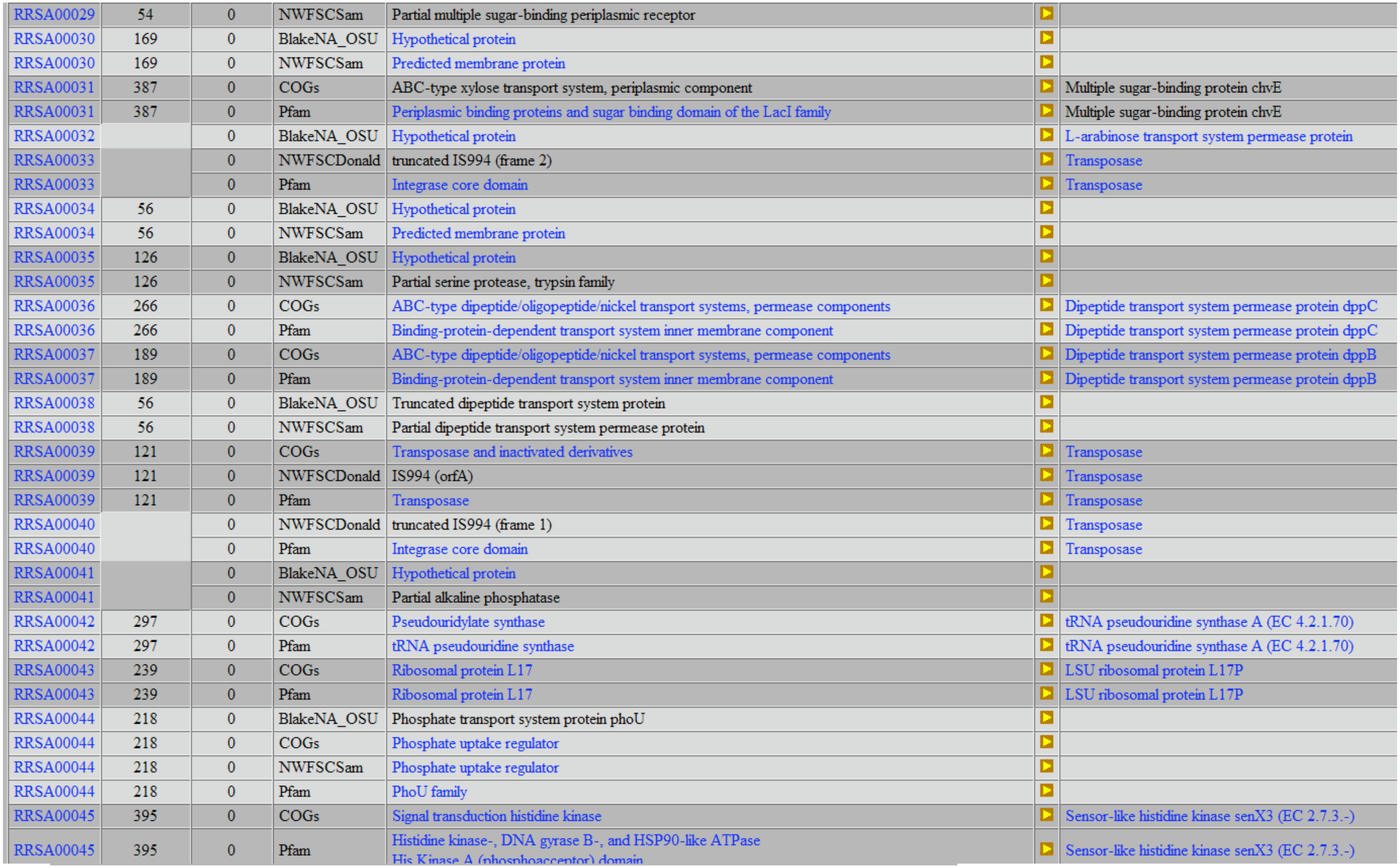
- 850 genomes in the database; 437 bacteria, 129 eukarya
- 2.2 million genes, >60% with detailed annotations
- Association of genes into >6300 metabolic pathways
- Interpretation of microarray data in metabolic context
- Comparative genomics approach: identifies more genes and functio

Manual curation generates highest quality gene annotations

Functional gene assignments: IG vs TIGR

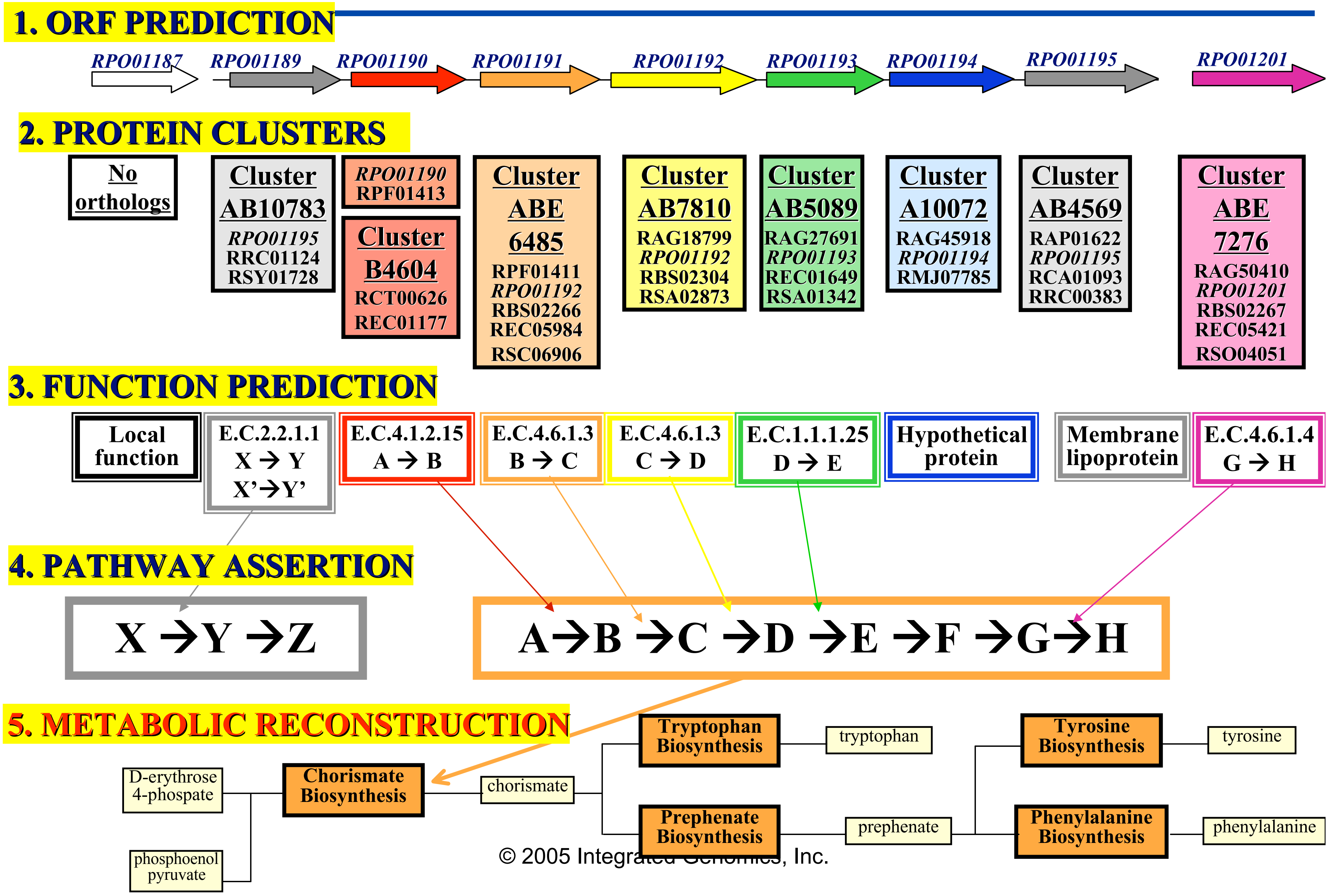
ORGANISM	# Genes	% IG	% TIGR
Brucella suis	3264	70	53
Chlamydia pneumonia AR39	1136	59	43
Chlamydia trachomatis MoPn	928	70	50
Haemophilus influenzae	1846	78	59
Mycobacterium tuberculosis	4473	61	42
Mycoplasma genitalium	532	71	68
Neisseria meningitidis MC58	2329	65	48
Pseudomonas putida	5350	70	65
Streptococcus pneumoniae	2304	73	53
Vibrio cholerae N16961	3915	66	51
Overall	**-**	**68%**	**53%**

ERGO™: compare annotations



RRSA00029	54	0	NWFSCSam	Partial multiple sugar-binding periplasmic receptor	▶	
RRSA00030	169	0	BlakeNA_OSU	Hypothetical protein	▶	
RRSA00030	169	0	NWFSCSam	Predicted membrane protein	▶	
RRSA00031	387	0	COGs	ABC-type xylose transport system, periplasmic component	▶	Multiple sugar-binding protein chvE
RRSA00031	387	0	Pfam	Periplasmic binding proteins and sugar binding domain of the LacI family	▶	Multiple sugar-binding protein chvE
RRSA00032		0	BlakeNA_OSU	Hypothetical protein	▶	L-arabinose transport system permease protein
RRSA00033		0	NWFSCDonald	truncated IS994 (frame 2)	▶	Transposase
RRSA00033		0	Pfam	Integrase core domain	▶	Transposase
RRSA00034	56	0	BlakeNA_OSU	Hypothetical protein	▶	
RRSA00034	56	0	NWFSCSam	Predicted membrane protein	▶	
RRSA00035	126	0	BlakeNA_OSU	Hypothetical protein	▶	
RRSA00035	126	0	NWFSCSam	Partial serine protease, trypsin family	▶	
RRSA00036	266	0	COGs	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components	▶	Dipeptide transport system permease protein dppC
RRSA00036	266	0	Pfam	Binding-protein-dependent transport system inner membrane component	▶	Dipeptide transport system permease protein dppC
RRSA00037	189	0	COGs	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components	▶	Dipeptide transport system permease protein dppB
RRSA00037	189	0	Pfam	Binding-protein-dependent transport system inner membrane component	▶	Dipeptide transport system permease protein dppB
RRSA00038	56	0	BlakeNA_OSU	Truncated dipeptide transport system protein	▶	
RRSA00038	56	0	NWFSCSam	Partial dipeptide transport system permease protein	▶	
RRSA00039	121	0	COGs	Transposase and inactivated derivatives	▶	Transposase
RRSA00039	121	0	NWFSCDonald	IS994 (orfA)	▶	Transposase
RRSA00039	121	0	Pfam	Transposase	▶	Transposase
RRSA00040		0	NWFSCDonald	truncated IS994 (frame 1)	▶	Transposase
RRSA00040		0	Pfam	Integrase core domain	▶	Transposase
RRSA00041		0	BlakeNA_OSU	Hypothetical protein	▶	
RRSA00041		0	NWFSCSam	Partial alkaline phosphatase	▶	
RRSA00042	297	0	COGs	Pseudouridylate synthase	▶	tRNA pseudouridine synthase A (EC 4.2.1.70)
RRSA00042	297	0	Pfam	tRNA pseudouridine synthase	▶	tRNA pseudouridine synthase A (EC 4.2.1.70)
RRSA00043	239	0	COGs	Ribosomal protein L17	▶	LSU ribosomal protein L17P
RRSA00043	239	0	Pfam	Ribosomal protein L17	▶	LSU ribosomal protein L17P
RRSA00044	218	0	BlakeNA_OSU	Phosphate transport system protein phoU	▶	
RRSA00044	218	0	COGs	Phosphate uptake regulator	▶	
RRSA00044	218	0	NWFSCSam	Phosphate uptake regulator	▶	
RRSA00044	218	0	Pfam	PhoU family	▶	
RRSA00045	395	0	COGs	Signal transduction histidine kinase	▶	Sensor-like histidine kinase senX3 (EC 2.7.3.-)
RRSA00045	395	0	Pfam	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase His Kinase A (phosphoacceptor) domain	▶	Sensor-like histidine kinase senX3 (EC 2.7.3.-)

ERGO™: Reconstruction steps



1. ORF PREDICTION

RPO01187 *RPO01189* *RPO01190* *RPO01191* *RPO01192* *RPO01193* *RPO01194* *RPO01195* *RPO01201*

2. PROTEIN CLUSTERS

No orthologs

Cluster AB10783
RPO01195
RRC01124
RSY01728

RPO01190
RPF01413

Cluster B4604
RCT00626
REC01177

Cluster ABE 6485
RPF01411
RPO01192
RBS02266
REC05984
RSC06906

Cluster AB7810
RAG18799
RPO01192
RBS02304
RSA02873

Cluster AB5089
RAG27691
RPO01193
REC01649
RSA01342

Cluster A10072
RAG45918
RPO01194
RMJ07785

Cluster AB4569
RAP01622
RPO01195
RCA01093
RRC00383

Cluster ABE 7276
RAG50410
RPO01201
RBS02267
REC05421
RSO04051

3. FUNCTION PREDICTION

Local function

E.C.2.2.1.1
X → Y
X'→Y'

E.C.4.1.2.15
A → B

E.C.4.6.1.3
B → C

E.C.4.6.1.3
C → D

E.C.1.1.1.25
D → E

Hypothetical protein

Membrane lipoprotein

E.C.4.6.1.4
G → H

4. PATHWAY ASSERTION

X →Y →Z

A→B →C →D →E →F →G→H

5. METABOLIC RECONSTRUCTION

D-erythrose 4-phospate

phosphoenol pyruvate

Chorismate Biosynthesis

chorismate

Tryptophan Biosynthesis

tryptophan

Prephenate Biosynthesis

prephenate

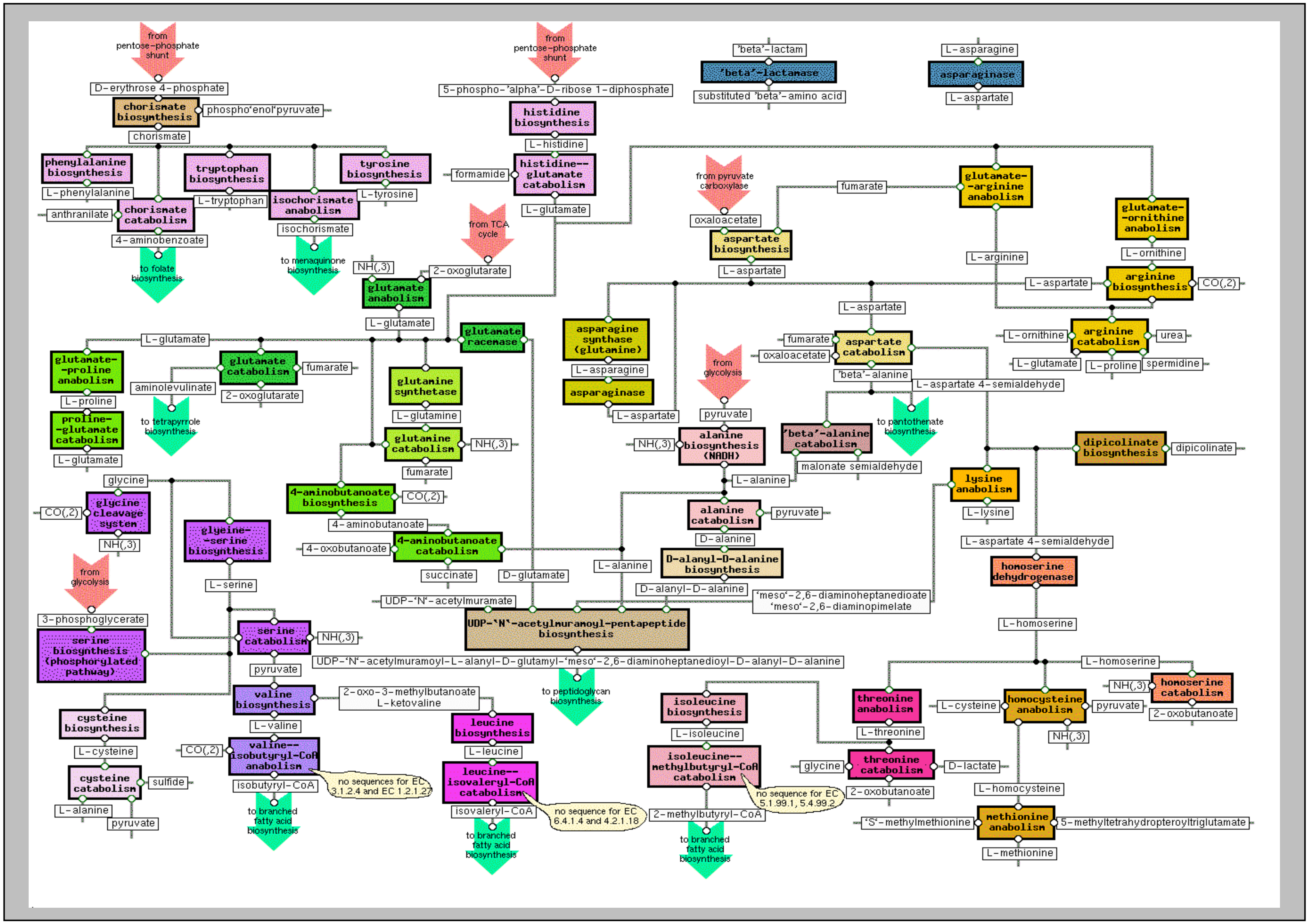
Tyrosine Biosynthesis

tyrosine

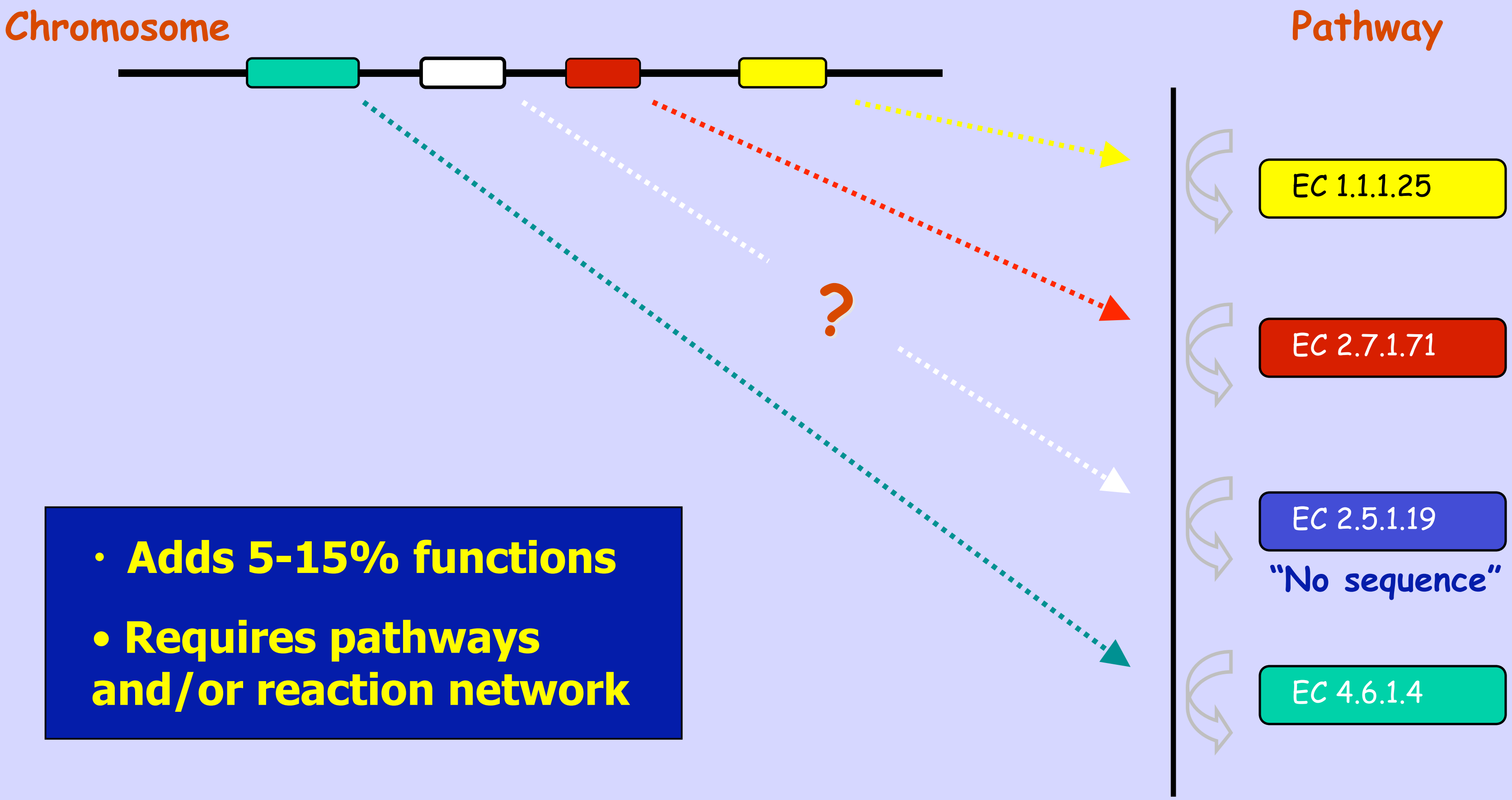
Phenylalanine Biosynthesis

phenylalanine

Functional Reconstruction: a ‘road-map’

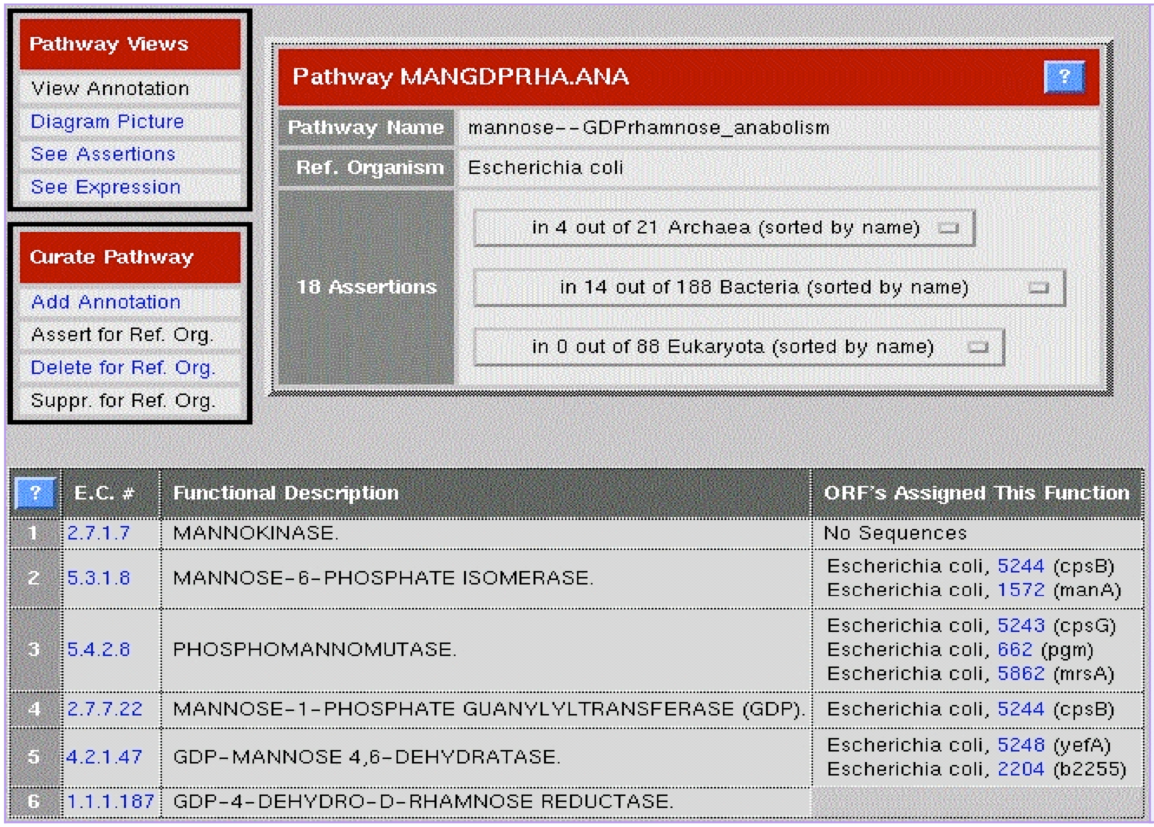


Pathway Mapping: identification of ‘missing genes’



- **Adds 5-15% functions**
- **Requires pathways and/or reaction network**

ERGO: Pathway Based Annotations



Pathway Views

- View Annotation
- Diagram Picture
- See Assertions
- See Expression

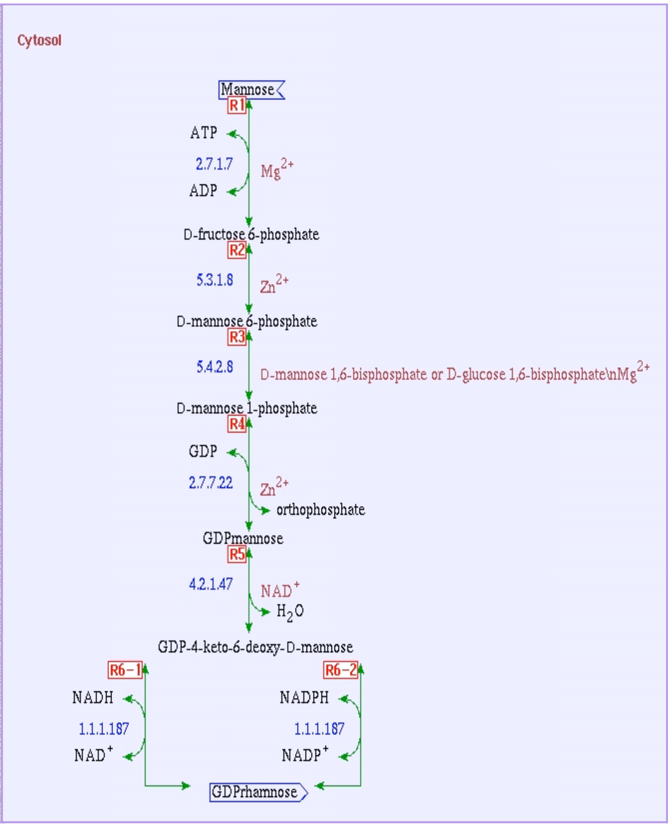
Curate Pathway

- Add Annotation
- Assert for Ref. Org.
- Delete for Ref. Org.
- Suppr. for Ref. Org.

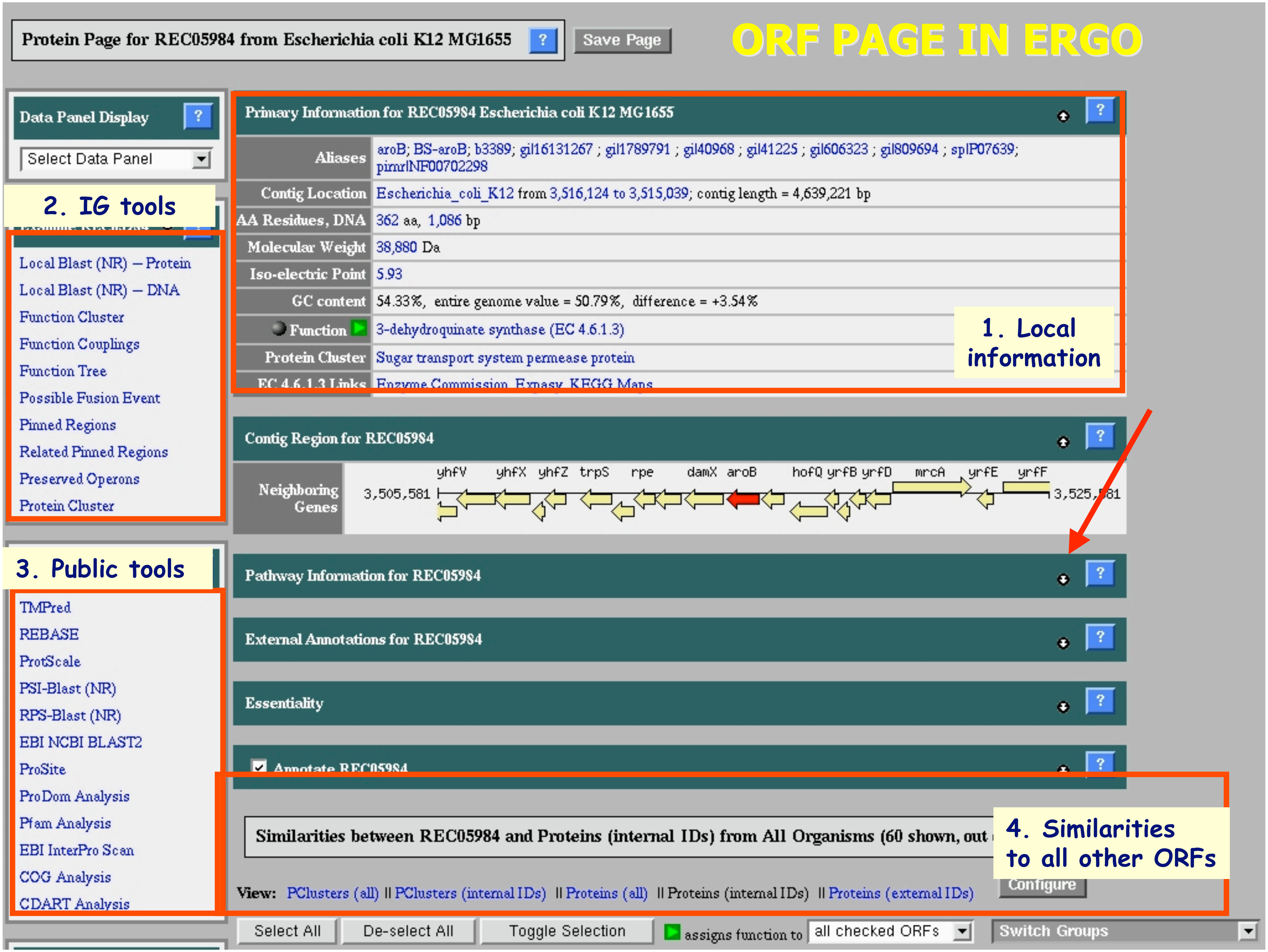
Pathway MANGDPRHA.ANA ?

Pathway Name	mannose--GDPrhamnose_anabolism
Ref. Organism	Escherichia coli
18 Assertions	in 4 out of 21 Archaea (sorted by name) in 14 out of 188 Bacteria (sorted by name) in 0 out of 88 Eukaryota (sorted by name)

?	E.C. #	Functional Description	ORF's Assigned This Function
1	2.7.1.7	MANNOKINASE.	No Sequences
2	5.3.1.8	MANNOSE-6-PHOSPHATE ISOMERASE.	Escherichia coli, 5244 (cpsB) Escherichia coli, 1572 (manA)
3	5.4.2.8	PHOSPHOMANNOMUTASE.	Escherichia coli, 5243 (cpsG) Escherichia coli, 662 (pgm) Escherichia coli, 5862 (mrsA)
4	2.7.7.22	MANNOSE-1-PHOSPHATE GUANYLYLTRANSFERASE (GDP).	Escherichia coli, 5244 (cpsB)
5	4.2.1.47	GDP-MANNOSE 4,6-DEHYDRATASE.	Escherichia coli, 5248 (yefA) Escherichia coli, 2204 (b2255)
6	1.1.1.187	GDP-4-DEHYDRO-D-RHAMNOSE REDUCTASE.	



ORF PAGE IN ERGO



Protein Page for REC05984 from Escherichia coli K12 MG1655 ? Save Page

Data Panel Display ?

Select Data Panel

2. IG tools

- Local Blast (NR) – Protein
- Local Blast (NR) – DNA
- Function Cluster
- Function Couplings
- Function Tree
- Possible Fusion Event
- Pinned Regions
- Related Pinned Regions
- Preserved Operons
- Protein Cluster

3. Public tools

- TMPred
- REBASE
- ProtScale
- PSI-Blast (NR)
- RPS-Blast (NR)
- EBI NCBI BLAST2
- ProSite
- ProDom Analysis
- Pfam Analysis
- EBI InterPro Scan
- COG Analysis
- CDART Analysis

Primary Information for REC05984 Escherichia coli K12 MG1655 ?

Aliases	aroB; BS-aroB; b3389; gil16131267 ; gil1789791 ; gil40968 ; gil41225 ; gil606323 ; gil809694 ; splP07639; pirnrlNF00702298
Contig Location	Escherichia_coli_K12 from 3,516,124 to 3,515,039; contig length = 4,639,221 bp
AA Residues, DNA	362 aa, 1,086 bp
Molecular Weight	38,880 Da
Iso-electric Point	5.93
GC content	54.33%, entire genome value = 50.79%, difference = +3.54%
● Function ▶	3-dehydroquinate synthase (EC 4.6.1.3)
Protein Cluster	Sugar transport system permease protein
EC 4.6.1.3 Links	Enzyme Commission Expasy KEGG Maps

1. Local information

Contig Region for REC05984 ?

Neighboring Genes

Pathway Information for REC05984 ?

External Annotations for REC05984 ?

Essentiality ?

☑ Annotate REC05984 ?

Similarities between REC05984 and Proteins (internal IDs) from All Organisms (60 shown, out

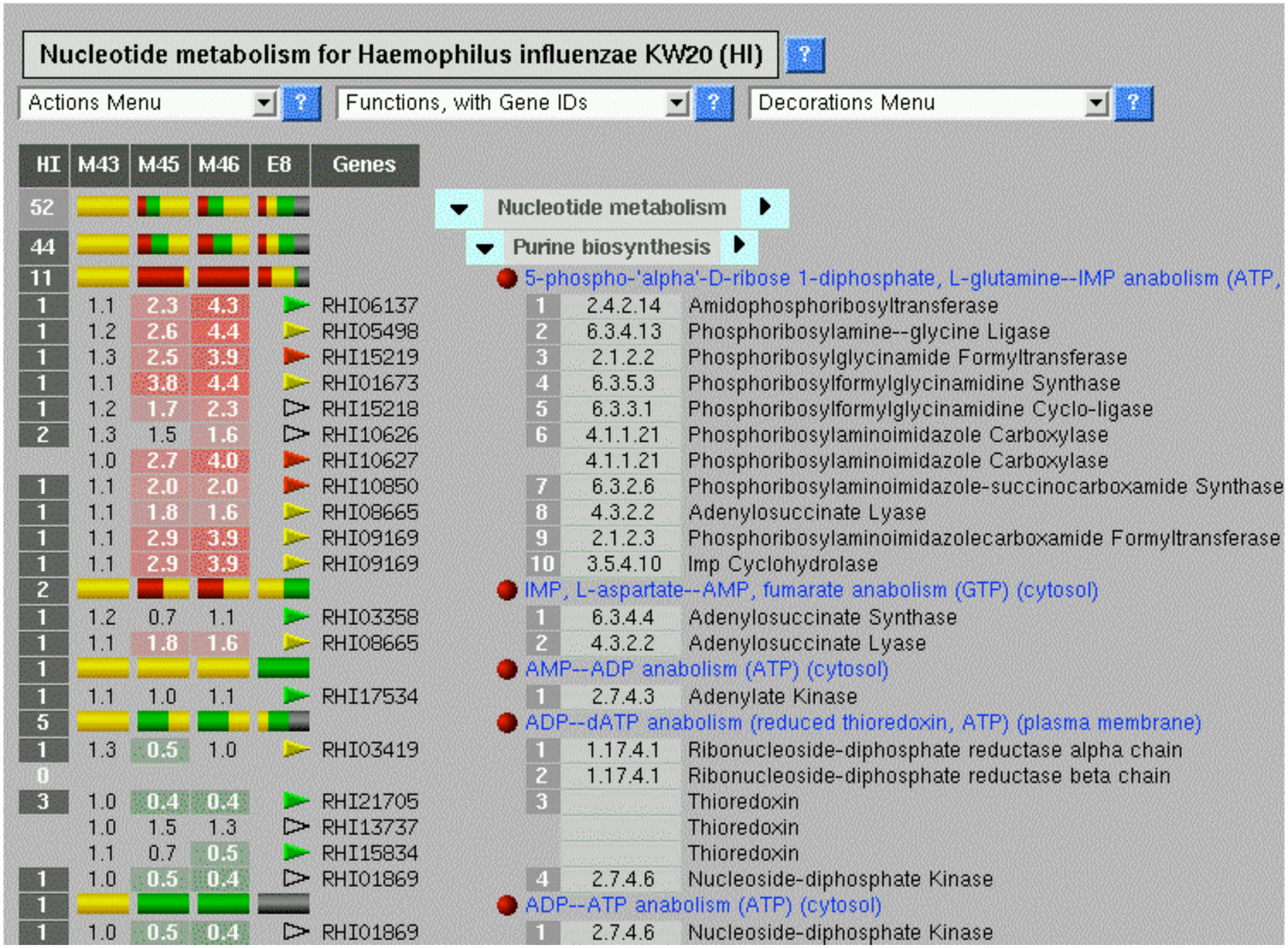
4. Similarities to all other ORFs

View: PClusters (all) || PClusters (internal IDs) || Proteins (all) || Proteins (internal IDs) || Proteins (external IDs)

Configure

Select All De-select All Toggle Selection ▶ assigns function to all checked ORFs Switch Groups

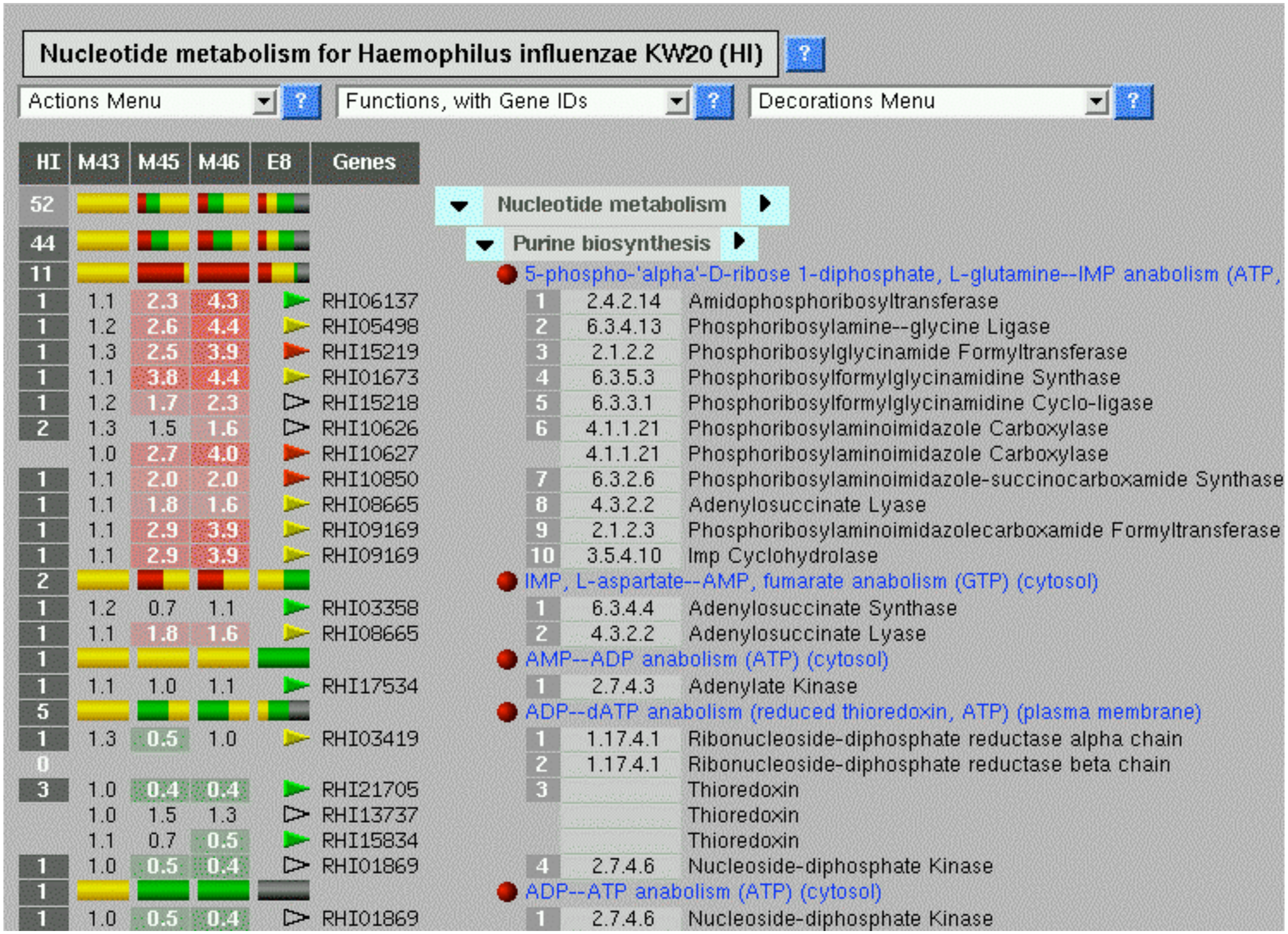
Integration of Gene Expression Data with Functional Overviews



Nucleotide metabolism for Haemophilus influenzae KW20 (HI)

HI	M43	M45	M46	E8	Genes			
52	🟨	🟥🟩🟨	🟥🟩🟨	🟥🟨🟩⬛		▼ Nucleotide metabolism ▶		
44	🟨	🟥🟩🟨	🟥🟩🟨	🟥🟨🟩⬛		▼ Purine biosynthesis ▶		
11	🟨	🟥	🟥	🟥🟨🟩		● 5-phospho-'alpha'-D-ribose 1-diphosphate, L-glutamine--IMP anabolism (ATP,		
1	1.1	2.3	4.3	▶	RHI06137	1	2.4.2.14	Amidophosphoribosyltransferase
1	1.2	2.6	4.4	▶	RHI05498	2	6.3.4.13	Phosphoribosylamine--glycine Ligase
1	1.3	2.5	3.9	▶	RHI15219	3	2.1.2.2	Phosphoribosylglycinamide Formyltransferase
1	1.1	3.8	4.4	▶	RHI01673	4	6.3.5.3	Phosphoribosylformylglycinamidine Synthase
1	1.2	1.7	2.3	▷	RHI15218	5	6.3.3.1	Phosphoribosylformylglycinamidine Cyclo-ligase
2	1.3	1.5	1.6	▷	RHI10626	6	4.1.1.21	Phosphoribosylaminoimidazole Carboxylase
	1.0	2.7	4.0	▶	RHI10627		4.1.1.21	Phosphoribosylaminoimidazole Carboxylase
1	1.1	2.0	2.0	▶	RHI10850	7	6.3.2.6	Phosphoribosylaminoimidazole-succinocarboxamide Synthase
1	1.1	1.8	1.6	▶	RHI08665	8	4.3.2.2	Adenylosuccinate Lyase
1	1.1	2.9	3.9	▶	RHI09169	9	2.1.2.3	Phosphoribosylaminoimidazolecarboxamide Formyltransferase
1	1.1	2.9	3.9	▶	RHI09169	10	3.5.4.10	Imp Cyclohydrolase
2	🟨	🟥🟨	🟥🟨	🟨🟩		● IMP, L-aspartate--AMP, fumarate anabolism (GTP) (cytosol)		
1	1.2	0.7	1.1	▶	RHI03358	1	6.3.4.4	Adenylosuccinate Synthase
1	1.1	1.8	1.6	▶	RHI08665	2	4.3.2.2	Adenylosuccinate Lyase
1	🟨	🟨	🟨	🟩		● AMP--ADP anabolism (ATP) (cytosol)		
1	1.1	1.0	1.1	▶	RHI17534	1	2.7.4.3	Adenylate Kinase
5	🟨	🟩🟨	🟩🟨	🟨🟩⬛		● ADP--dATP anabolism (reduced thioredoxin, ATP) (plasma membrane)		
1	1.3	0.5	1.0	▶	RHI03419	1	1.17.4.1	Ribonucleoside-diphosphate reductase alpha chain
0						2	1.17.4.1	Ribonucleoside-diphosphate reductase beta chain
3	1.0	0.4	0.4	▶	RHI21705	3		Thioredoxin
	1.0	1.5	1.3	▷	RHI13737			Thioredoxin
	1.1	0.7	0.5	▶	RHI15834			Thioredoxin
1	1.0	0.5	0.4	▷	RHI01869	4	2.7.4.6	Nucleoside-diphosphate Kinase
1	🟨	🟩	🟩	⬛		● ADP--ATP anabolism (ATP) (cytosol)		
1	1.0	0.5	0.4	▷	RHI01869	1	2.7.4.6	Nucleoside-diphosphate Kinase

Integration of Essentiality Data on Functional Overviews



Nucleotide metabolism for Haemophilus influenzae KW20 (HI)

HI	M43	M45	M46	E8	Genes			
52								Nucleotide metabolism
44								Purine biosynthesis
11								● 5-phospho-'alpha'-D-ribose 1-diphosphate, L-glutamine--IMP anabolism (ATP,
1	1.1	2.3	4.3	▶	RHI06137	1	2.4.2.14	Amidophosphoribosyltransferase
1	1.2	2.6	4.4	▶	RHI05498	2	6.3.4.13	Phosphoribosylamine--glycine Ligase
1	1.3	2.5	3.9	▶	RHI15219	3	2.1.2.2	Phosphoribosylglycinamide Formyltransferase
1	1.1	3.8	4.4	▶	RHI01673	4	6.3.5.3	Phosphoribosylformylglycinamidine Synthase
1	1.2	1.7	2.3	▷	RHI15218	5	6.3.3.1	Phosphoribosylformylglycinamidine Cyclo-ligase
2	1.3	1.5	1.6	▷	RHI10626	6	4.1.1.21	Phosphoribosylaminoimidazole Carboxylase
	1.0	2.7	4.0	▶	RHI10627		4.1.1.21	Phosphoribosylaminoimidazole Carboxylase
1	1.1	2.0	2.0	▶	RHI10850	7	6.3.2.6	Phosphoribosylaminoimidazole-succinocarboxamide Synthase
1	1.1	1.8	1.6	▶	RHI08665	8	4.3.2.2	Adenylosuccinate Lyase
1	1.1	2.9	3.9	▶	RHI09169	9	2.1.2.3	Phosphoribosylaminoimidazolecarboxamide Formyltransferase
1	1.1	2.9	3.9	▶	RHI09169	10	3.5.4.10	Imp Cyclohydrolase
2								● IMP, L-aspartate--AMP, fumarate anabolism (GTP) (cytosol)
1	1.2	0.7	1.1	▶	RHI03358	1	6.3.4.4	Adenylosuccinate Synthase
1	1.1	1.8	1.6	▶	RHI08665	2	4.3.2.2	Adenylosuccinate Lyase
1								● AMP--ADP anabolism (ATP) (cytosol)
1	1.1	1.0	1.1	▶	RHI17534	1	2.7.4.3	Adenylate Kinase
5								● ADP--dATP anabolism (reduced thioredoxin, ATP) (plasma membrane)
1	1.3	0.5	1.0	▶	RHI03419	1	1.17.4.1	Ribonucleoside-diphosphate reductase alpha chain
0						2	1.17.4.1	Ribonucleoside-diphosphate reductase beta chain
3	1.0	0.4	0.4	▶	RHI21705	3		Thioredoxin
	1.0	1.5	1.3	▷	RHI13737			Thioredoxin
	1.1	0.7	0.5	▶	RHI15834			Thioredoxin
1	1.0	0.5	0.4	▷	RHI01869	4	2.7.4.6	Nucleoside-diphosphate Kinase
1								● ADP--ATP anabolism (ATP) (cytosol)
1	1.0	0.5	0.4	▷	RHI01869	1	2.7.4.6	Nucleoside-diphosphate Kinase

Case study 1: *Xylella fastidiosa*

Comparative genomics and metabolic reconstruction applied to strain analysis

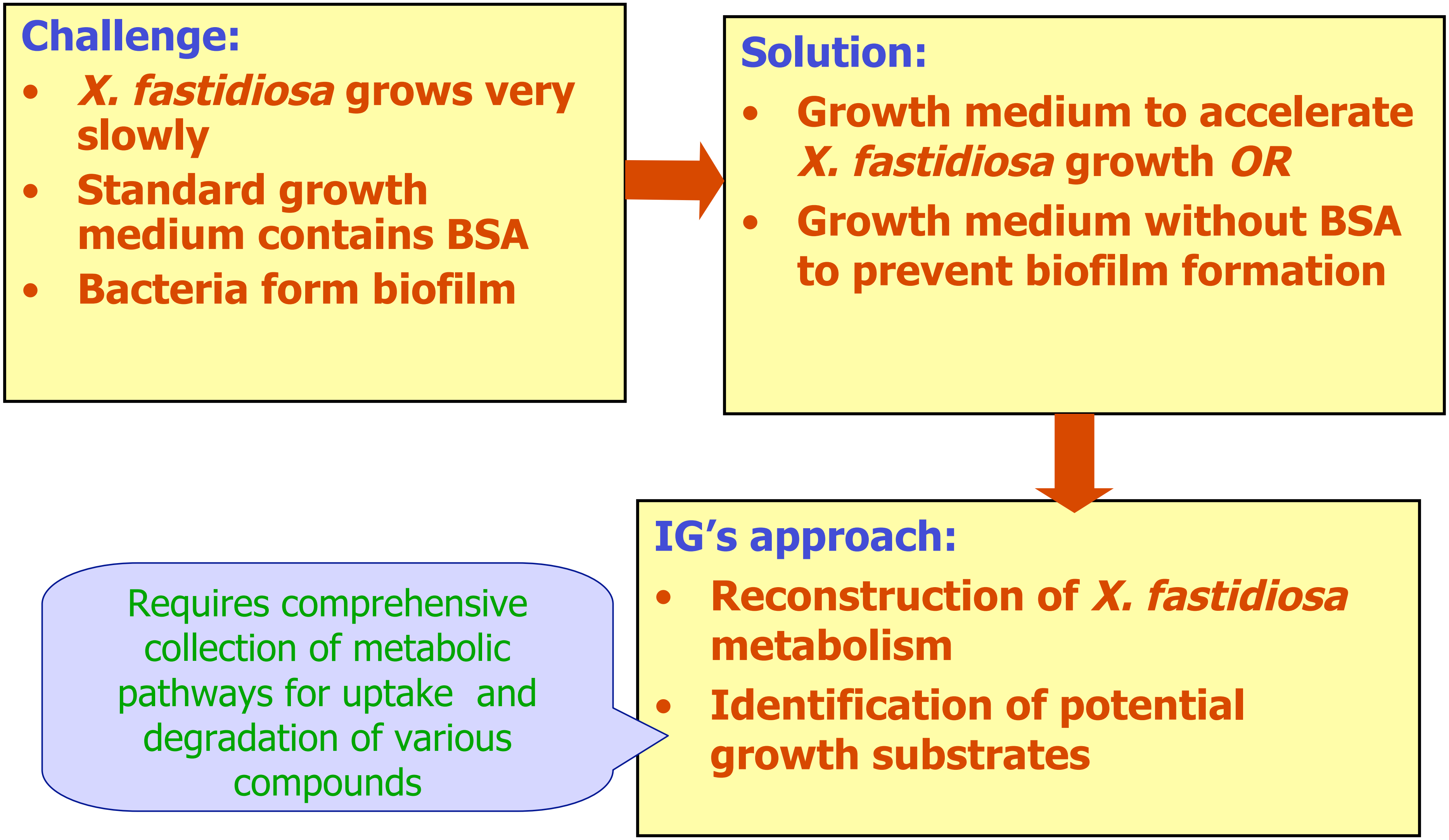
Collaborators:
U.S. Department of Energy (Joint Genome Institute)
University of California, Berkeley

References:
Bhattacharyya *et al* (2002) *Genome Research* 12:1556-1563
Bhattacharyya *et al* (2002) *Proc. Natl. Acad. Sci. USA* 99:12403-8

Xylella fastidiosa: background

- **Plant pathogens**
  - 3 strains sequenced, public genomes
  - *Xf* pv. citrus, *Xf* pv. almond, *Xf* pv. oleander
  - *Xf* pv. citrus (complete),
  - *Xf* pv. almond, *Xf* pv. oleander (10x draft)
- **Annotation**
  - Manual annotation and database curation
- **Pathway assertion**
- **Genome comparison (3 organisms)**
- **Metabolic reconstruction**
  - Predict physiology based on metabolic potential

Prediction of Growth Medium Components for *Xylella fastidiosa*



Challenge:

- ***X. fastidiosa*** **grows very slowly**
- **Standard growth medium contains BSA**
- **Bacteria form biofilm**

Solution:

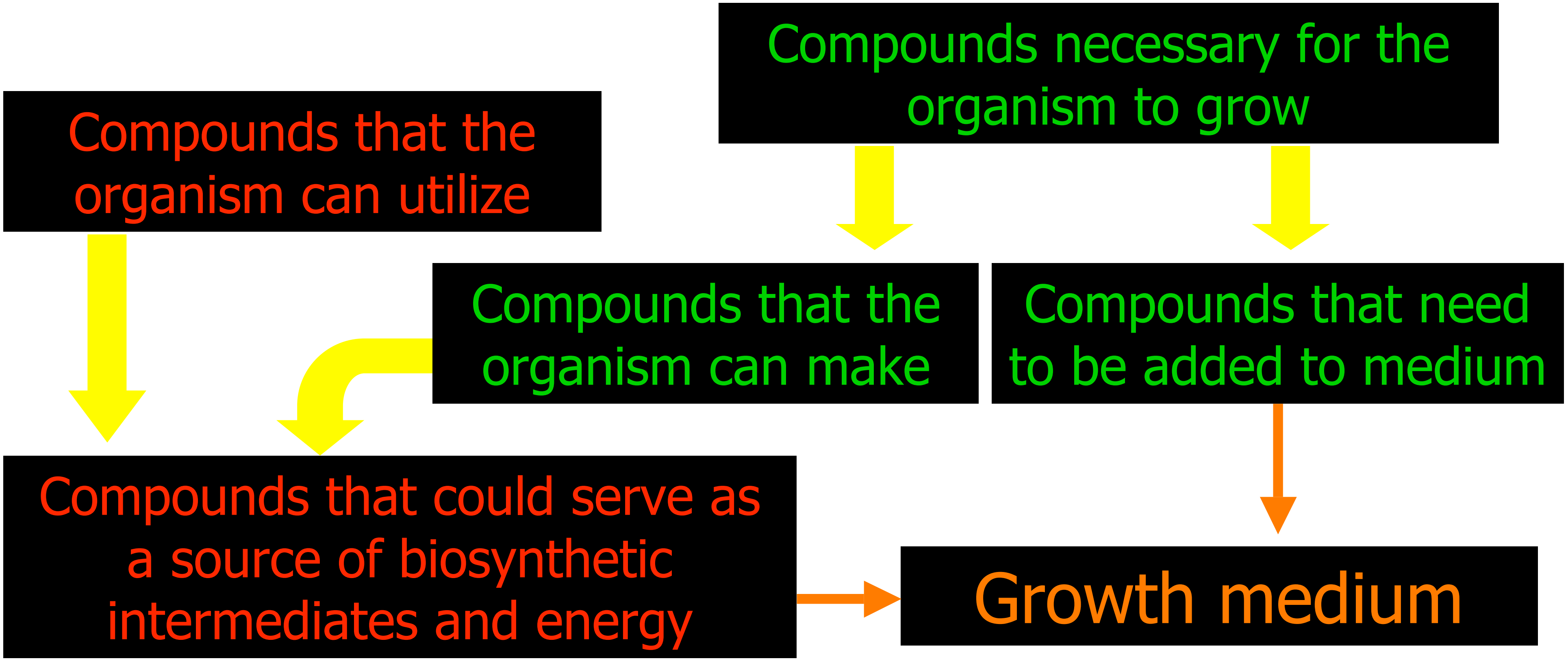
- **Growth medium to accelerate *X. fastidiosa* growth *OR***
- **Growth medium without BSA to prevent biofilm formation**

IG's approach:

- **Reconstruction of *X. fastidiosa* metabolism**
- **Identification of potential growth substrates**

Requires comprehensive collection of metabolic pathways for uptake and degradation of various compounds

Design of Growth Media: comparative genomics and functional reconstruction



Compounds that the organism can utilize

Compounds necessary for the organism to grow

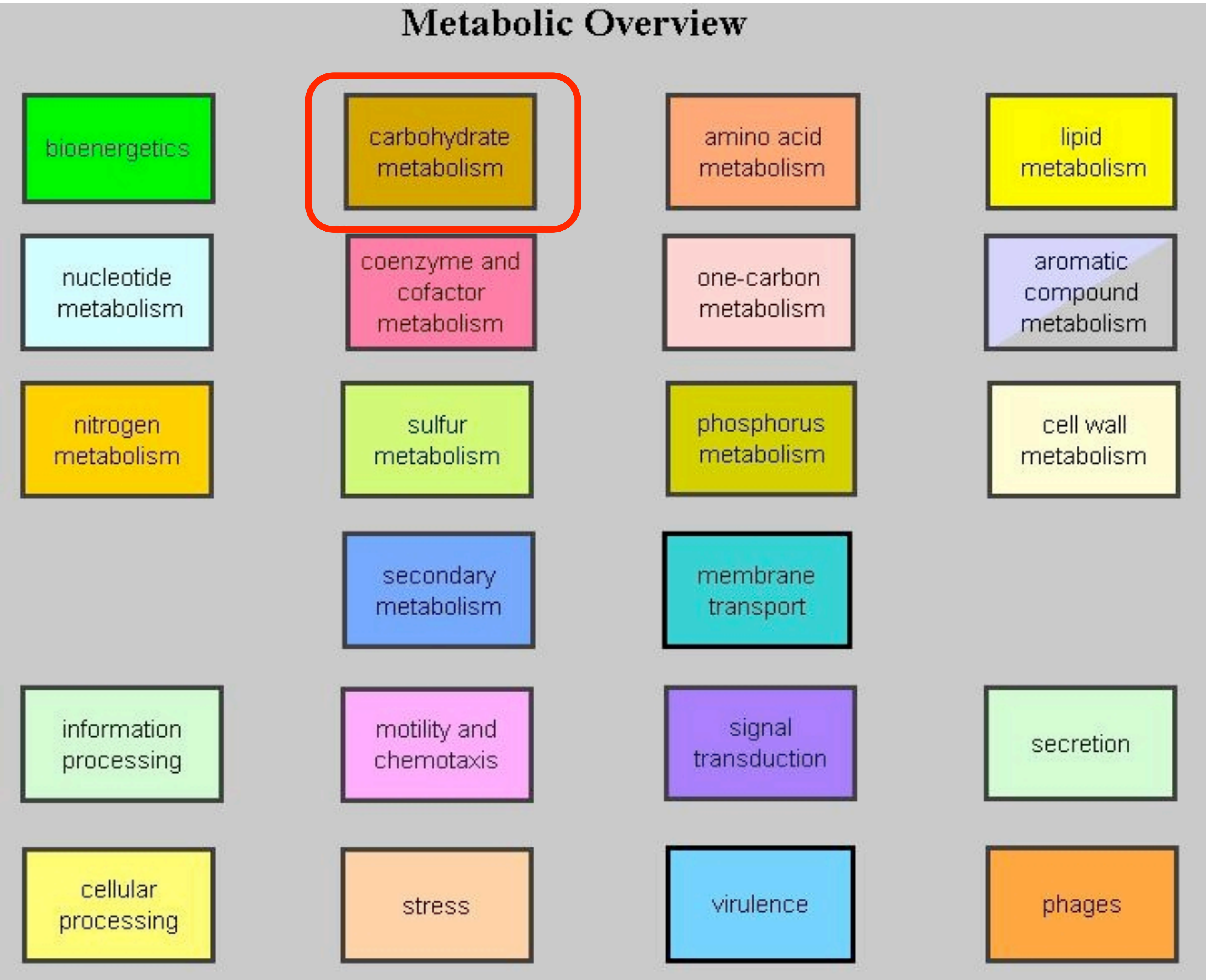
Compounds that the organism can make

Compounds that need to be added to medium

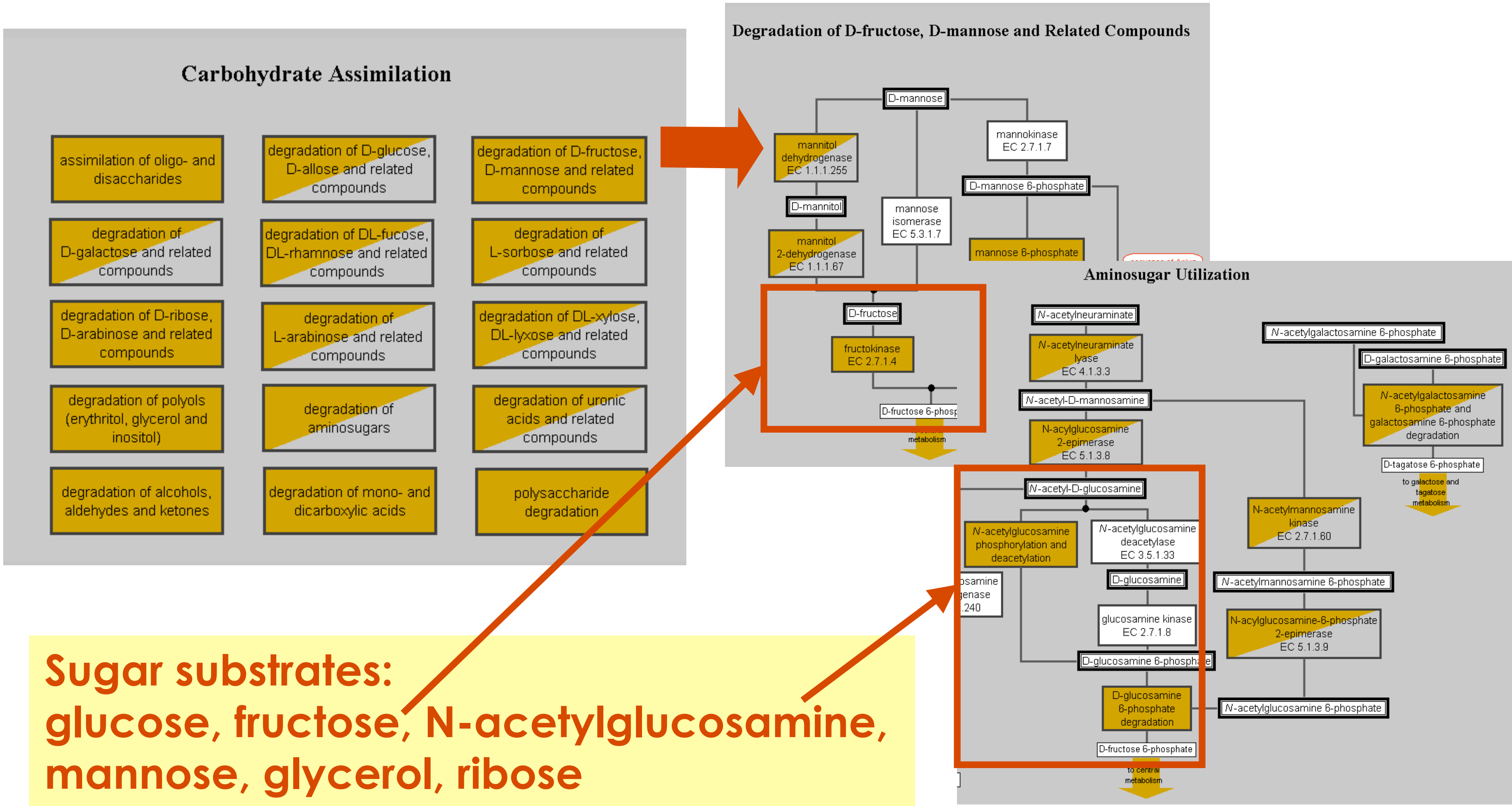
Compounds that could serve as a source of biosynthetic intermediates and energy

Growth medium

ERGO: Functional overview

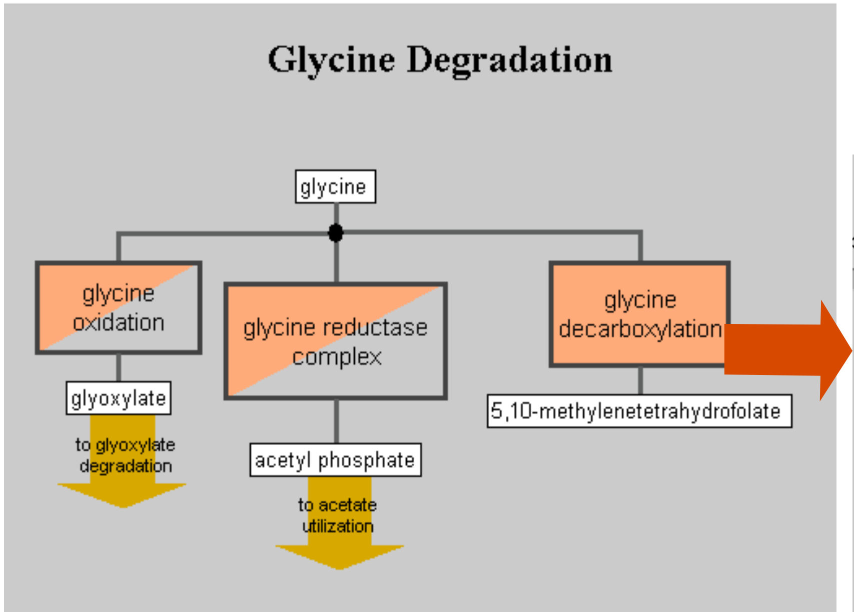


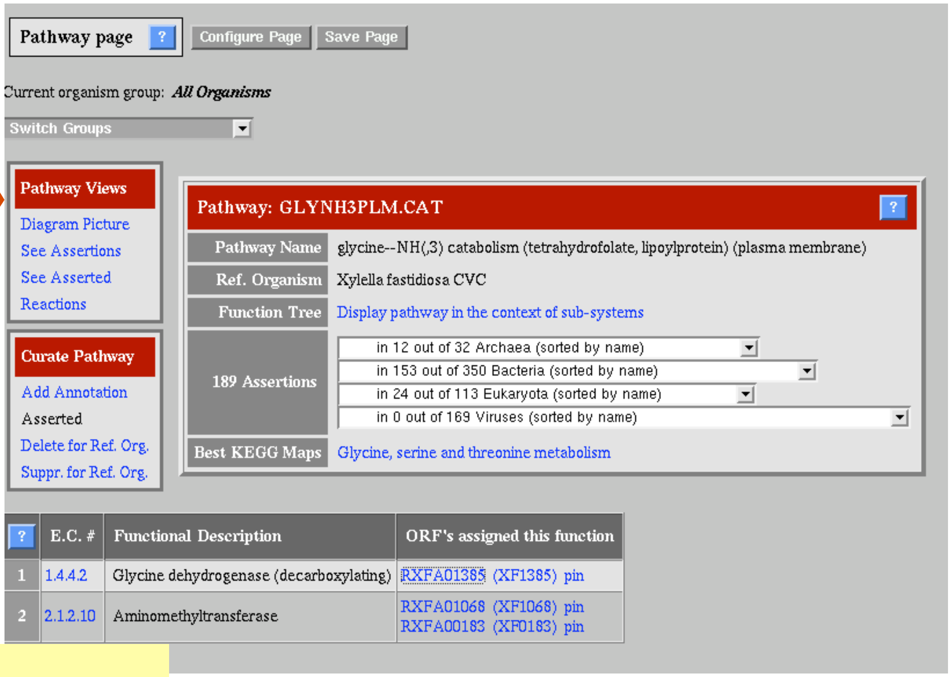
Analysis of potential growth substrates for *X. fastidiosa*


Carbohydrate Assimilation
assimilation of oligo- and disaccharides
degradation of D-glucose, D-allose and related compounds
degradation of D-fructose, D-mannose and related compounds
degradation of D-galactose and related compounds
degradation of DL-fucose, DL-rhamnose and related compounds
degradation of L-sorbose and related compounds
degradation of D-ribose, D-arabinose and related compounds
degradation of L-arabinose and related compounds
degradation of DL-xylose, DL-lyxose and related compounds
degradation of polyols (erythritol, glycerol and inositol)
degradation of aminosugars
degradation of uronic acids and related compounds
degradation of alcohols, aldehydes and ketones
degradation of mono- and dicarboxylic acids
polysaccharide degradation
Degradation of D-fructose, D-mannose and Related Compounds
D-mannose
mannitol dehydrogenase EC 1.1.1.255
mannokinase EC 2.7.1.7
D-mannitol
mannose isomerase EC 5.3.1.7
D-mannose 6-phosphate
mannitol 2-dehydrogenase EC 1.1.1.67
mannose 6-phosphate
D-fructose
fructokinase EC 2.7.1.4
D-fructose 6-phosp
metabolism
Aminosugar Utilization
N-acetylneuraminate
N-acetylneuraminate lyase EC 4.1.3.3
N-acetylgalactosamine 6-phosphate
D-galactosamine 6-phosphate
N-acetyl-D-mannosamine
N-acetylgalactosamine 6-phosphate and galactosamine 6-phosphate degradation
N-acylglucosamine 2-epimerase EC 5.1.3.8
D-tagatose 6-phosphate
to galactose and tagatose metabolism
N-acetyl-D-glucosamine
N-acetylmannosamine kinase EC 2.7.1.60
N-acetylglucosamine phosphorylation and deacetylation
N-acetylglucosamine deacetylase EC 3.5.1.33
osamine genase .240
D-glucosamine
N-acetylmannosamine 6-phosphate
glucosamine kinase EC 2.7.1.8
N-acylglucosamine-6-phosphate 2-epimerase EC 5.1.3.9
D-glucosamine 6-phosphate
D-glucosamine 6-phosphate degradation
N-acetylglucosamine 6-phosphate
D-fructose 6-phosphate
to central metabolism

Sugar substrates:
glucose, fructose, N-acetylglucosamine, mannose, glycerol, ribose

Identification of amino acid substrates for *X. fastidiosa*





Current organism group: ***All Organisms***

Pathway: GLYNH3PLM.CAT

Pathway Name	glycine--NH(,3) catabolism (tetrahydrofolate, lipoylprotein) (plasma membrane)
Ref. Organism	Xylella fastidiosa CVC
Function Tree	Display pathway in the context of sub-systems
189 Assertions	in 12 out of 32 Archaea (sorted by name) in 153 out of 350 Bacteria (sorted by name) in 24 out of 113 Eukaryota (sorted by name) in 0 out of 169 Viruses (sorted by name)
Best KEGG Maps	Glycine, serine and threonine metabolism

?	E.C. #	Functional Description	ORF's assigned this function
1	1.4.4.2	Glycine dehydrogenase (decarboxylating)	RXFA01385 (XF1385) pin
2	2.1.2.10	Aminomethyltransferase	RXFA01068 (XF1068) pin RXFA00183 (XF0183) pin

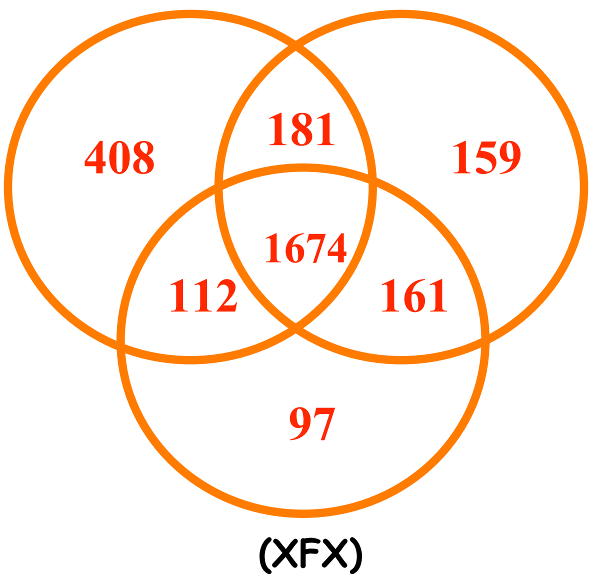
Amino acid substrates:
glycine, L-glutamate (2-ketoglutarate),
D- and L-alanine

Result: successful design of growth medium for *X. fastidiosa*

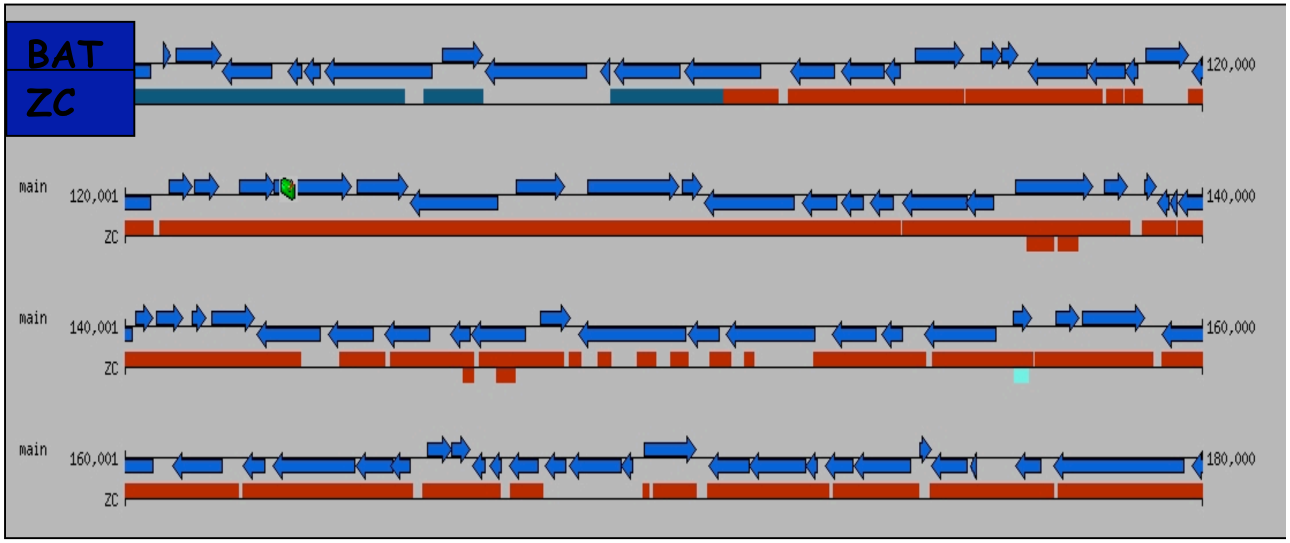
Medium	Growth, CFU
Standard PW (contains BSA)	1.13 x 10^9
Standard PW - BSA	no growth
Standard PW - BSA + fructose	8.4 x 10^9
Standard PW - BSA + fructose + glycine + L-alanine	9.5 x 10^9

Genome Comparisons: tools

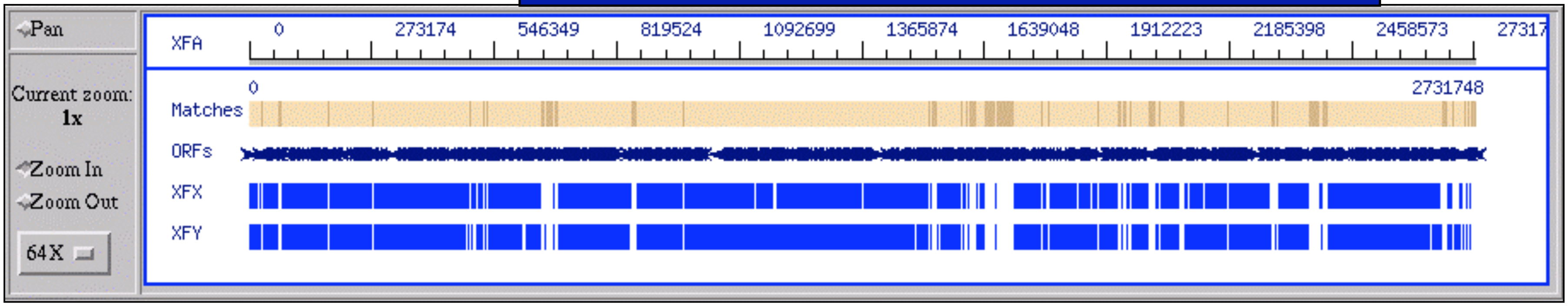
WorkBench: protein clustering



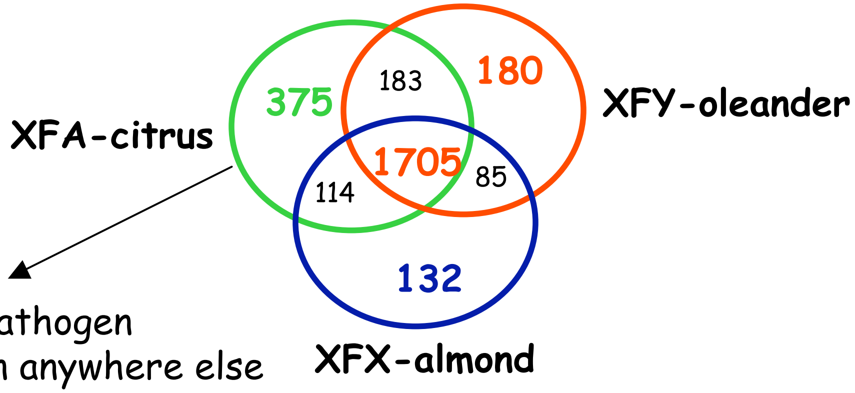
GenomeAlign: genomic DNA alignment



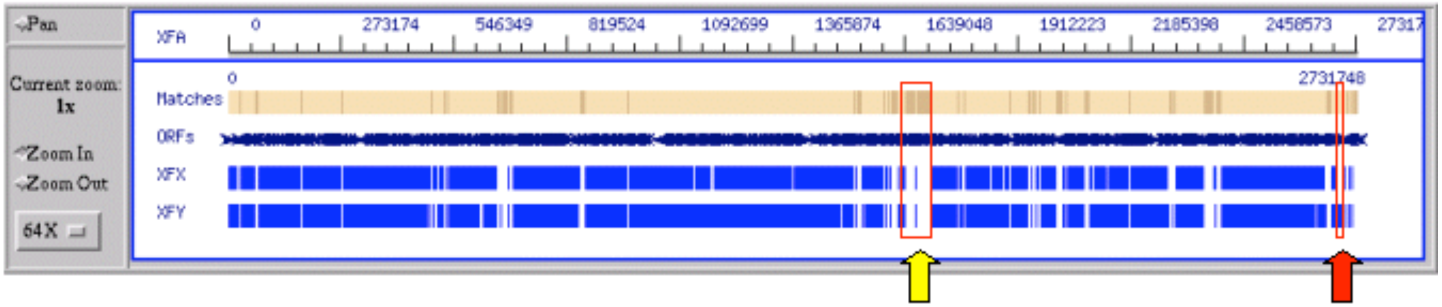
GenomeWalk: proteome alignment



Strain Comparisons: Protein clusters



Unique for the citrus pathogen
Of these, 54% not seen anywhere else
Includes a phage insertion



Phage insertion with 2 unusual carbon utilization operons conferring host specificity to grow in citrus

Case study 2: *Fusobacterium nucleatum*

Identification of genetic determinants of phenotypic traits (bad oral odor)

References:
Kapatral *et al* (2002) *J. Bact. 184: 2005-18*

F. nucleatum is a "BRIDGE" bacterium

- Over 300 genera and 500 species co-exist in oral cavity
  - Most are commensals but a few are opportunistic pathogens
- Infection process:
  - a) Tooth surface allows pellicle formation
  - b) Early colonizers: *Streptococci, Actinomyces spp*
  - c) *Fusobacteria spp.*
  - d) Late colonizers include pathogens:
    - *P. gingivalis, A. actinomycetemcomitans*
    - *T. denticola, B. forsythus*
- What is the physiological basis of mal-odour during infection?

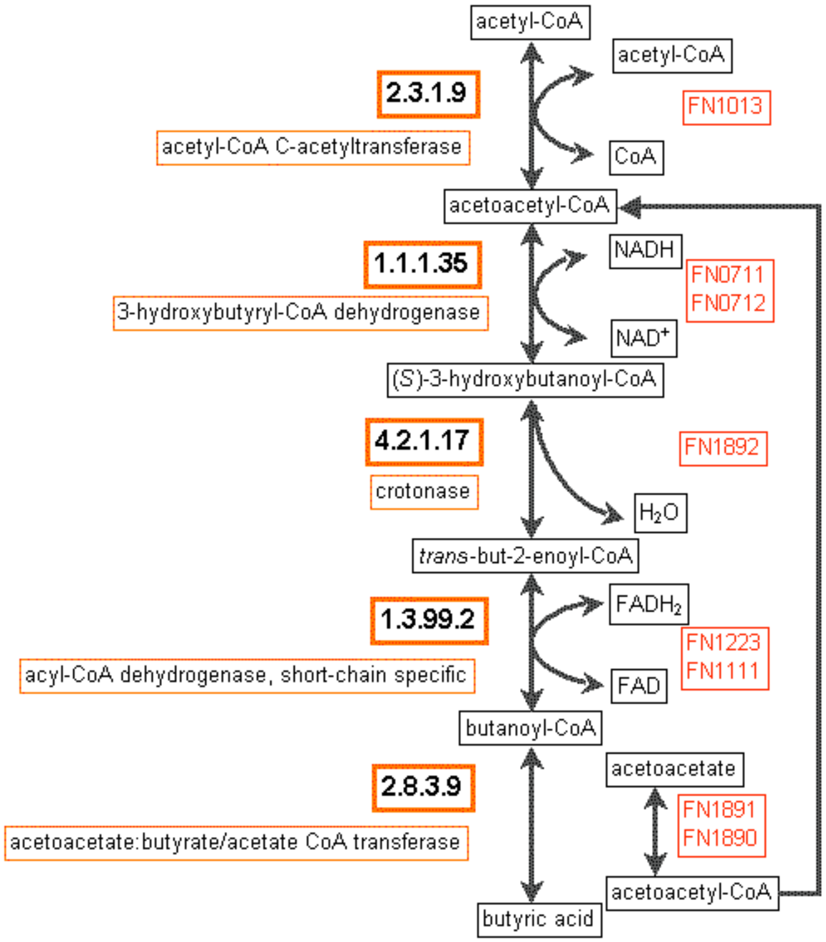
Species	Disease
F. nucleatum	Periodontitis
F. necrophorum	Lemierre's Syndrome
F. ulcercans	Skin ulcers
F. russi	Animal bites
F. varium	Eye infections

Fusobacterium: project schema

- **Determine genome sequence**
  - Whole genome shotgun sequencing
  - Primer walking to close gaps
  - Assembly
- **Annotation**
  - Automated ORF calling
  - Manual annotation and database curation
  - Identification of operons
- **Pathway assertion**
- **Metabolic reconstruction**
  - Predict physiology based on metabolic potential

What causes the bad smell ?

- Hydrogen sulfide: degradation of cysteine and methionine
- Butyric acid: prevents wound healing

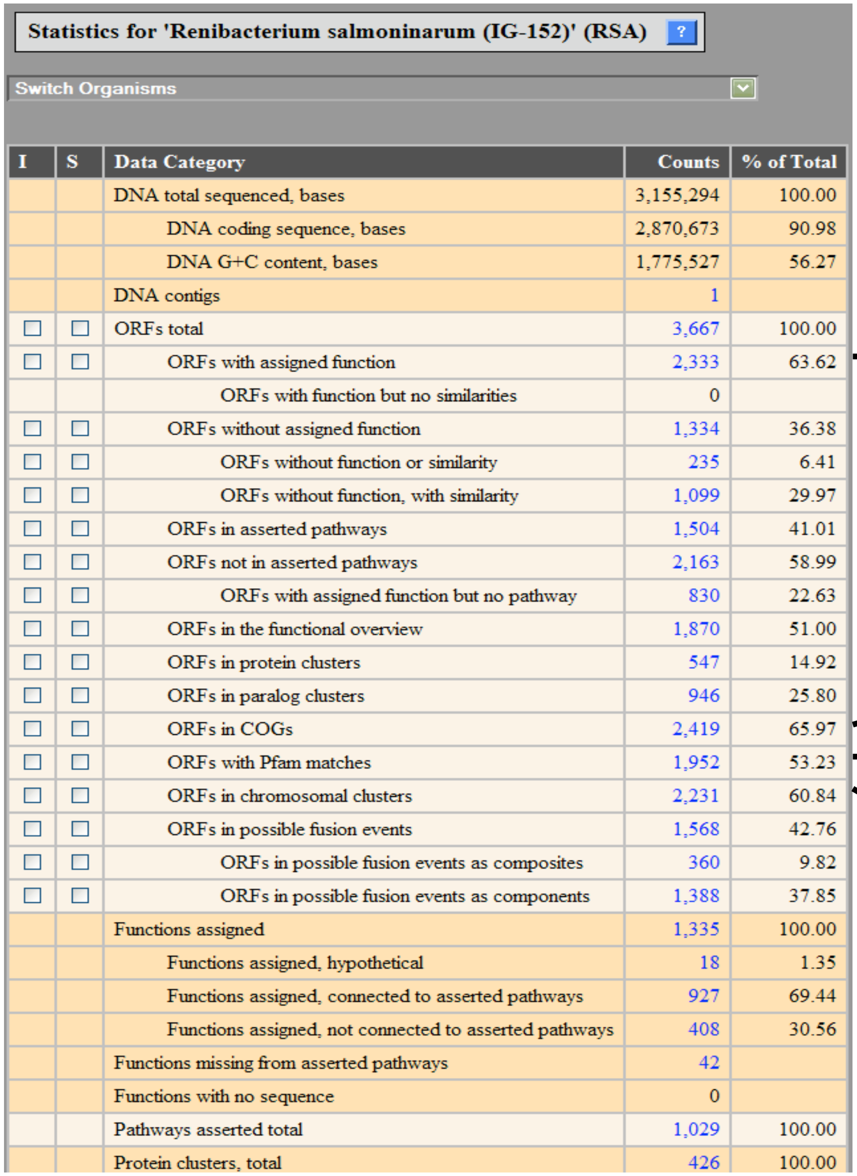


Reconstruction of the entire butyric acid biosynthetic pathway

BKD: Renibacterium salmoninarum

Genome Analysis and Functional Reconstruction: initial observations

Renibacterium: statistics



Statistics for 'Renibacterium salmoninarum (IG-152)' (RSA) ?

I	S	Data Category	Counts	% of Total
		DNA total sequenced, bases	3,155,294	100.00
		DNA coding sequence, bases	2,870,673	90.98
		DNA G+C content, bases	1,775,527	56.27
		DNA contigs	1	
□	□	ORFs total	3,667	100.00
□	□	ORFs with assigned function	2,333	63.62
		ORFs with function but no similarities	0	
□	□	ORFs without assigned function	1,334	36.38
□	□	ORFs without function or similarity	235	6.41
□	□	ORFs without function, with similarity	1,099	29.97
□	□	ORFs in asserted pathways	1,504	41.01
□	□	ORFs not in asserted pathways	2,163	58.99
□	□	ORFs with assigned function but no pathway	830	22.63
□	□	ORFs in the functional overview	1,870	51.00
□	□	ORFs in protein clusters	547	14.92
□	□	ORFs in paralog clusters	946	25.80
□	□	ORFs in COGs	2,419	65.97
□	□	ORFs with Pfam matches	1,952	53.23
□	□	ORFs in chromosomal clusters	2,231	60.84
□	□	ORFs in possible fusion events	1,568	42.76
□	□	ORFs in possible fusion events as composites	360	9.82
□	□	ORFs in possible fusion events as components	1,388	37.85
		Functions assigned	1,335	100.00
		Functions assigned, hypothetical	18	1.35
		Functions assigned, connected to asserted pathways	927	69.44
		Functions assigned, not connected to asserted pathways	408	30.56
		Functions missing from asserted pathways	42	
		Functions with no sequence	0	
		Pathways asserted total	1,029	100.00
		Protein clusters, total	426	100.00

After automated assignments

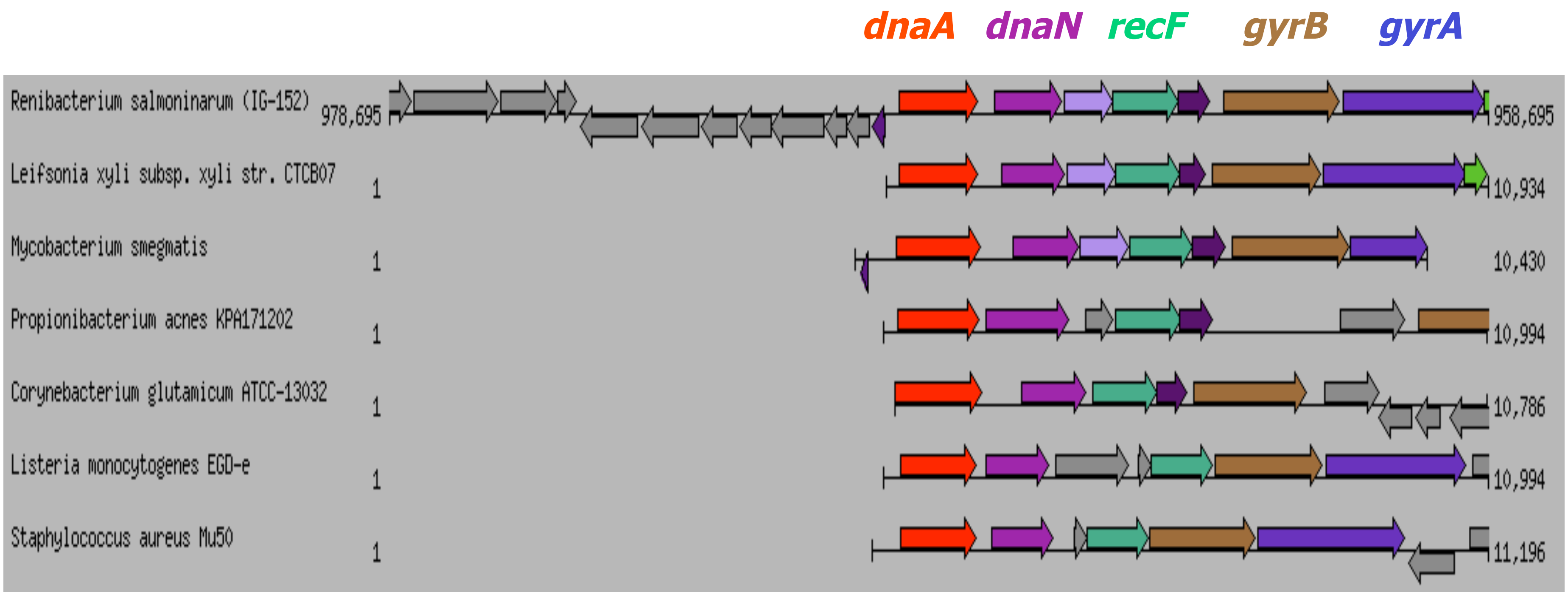
64% assigned functions

66% COGs families

53% Pfam domains

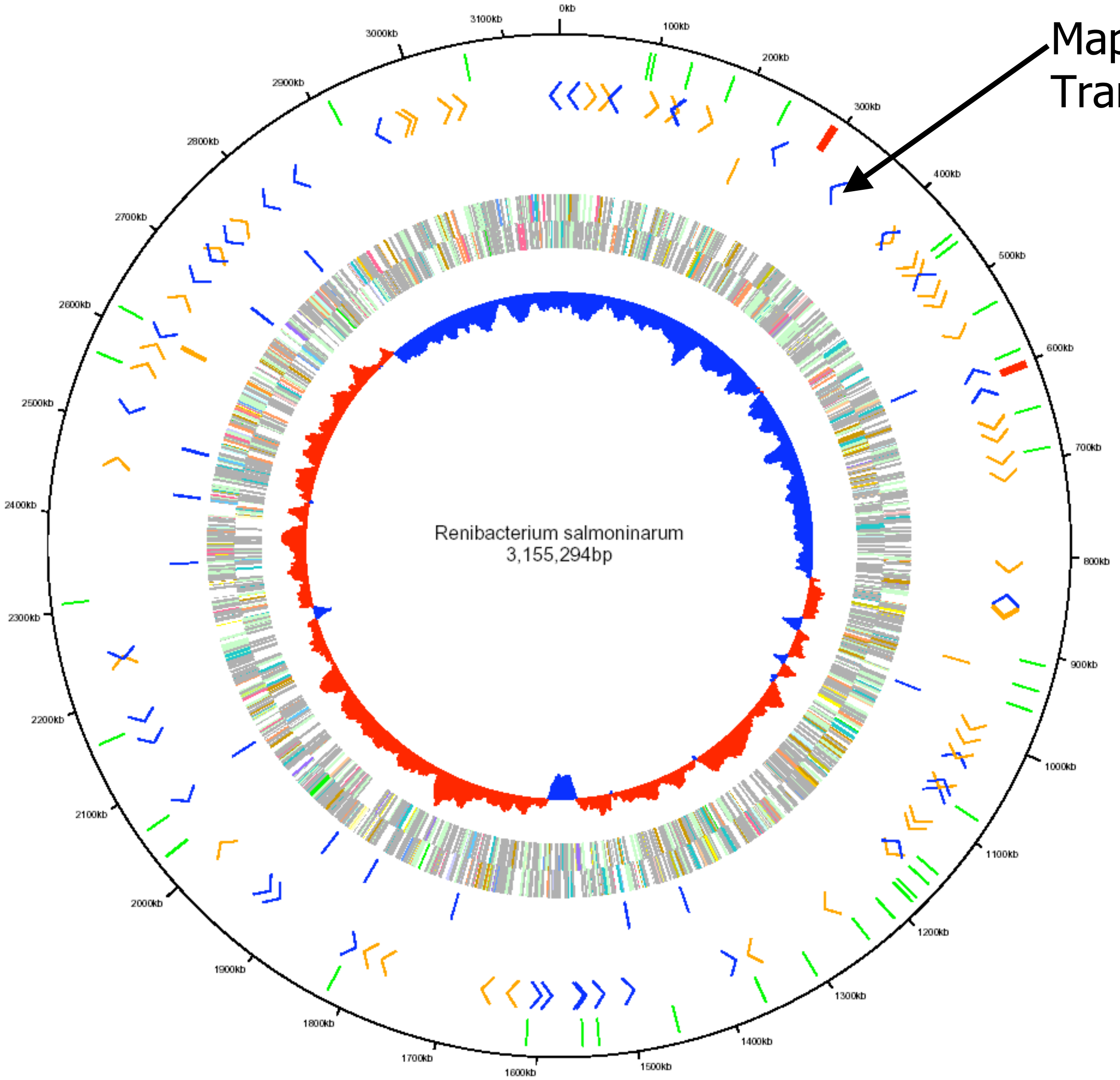
61% clustered ('operons')

Renibacterium: replication origin?


dnaA
dnaN
recF
gyrB
gyrA
Renibacterium salmoninarum (IG-152)
978,695
958,695
Leifsonia xyli subsp. xyli str. CTCB07
1
10,934
Mycobacterium smegmatis
1
10,430
Propionibacterium acnes KPA171202
1
10,994
Corynebacterium glutamicum ATCC-13032
1
10,786
Listeria monocytogenes EGD-e
1
10,994
Staphylococcus aureus Mu50
1
11,196

dnaA coordinates: begin 969,395; end 967,995

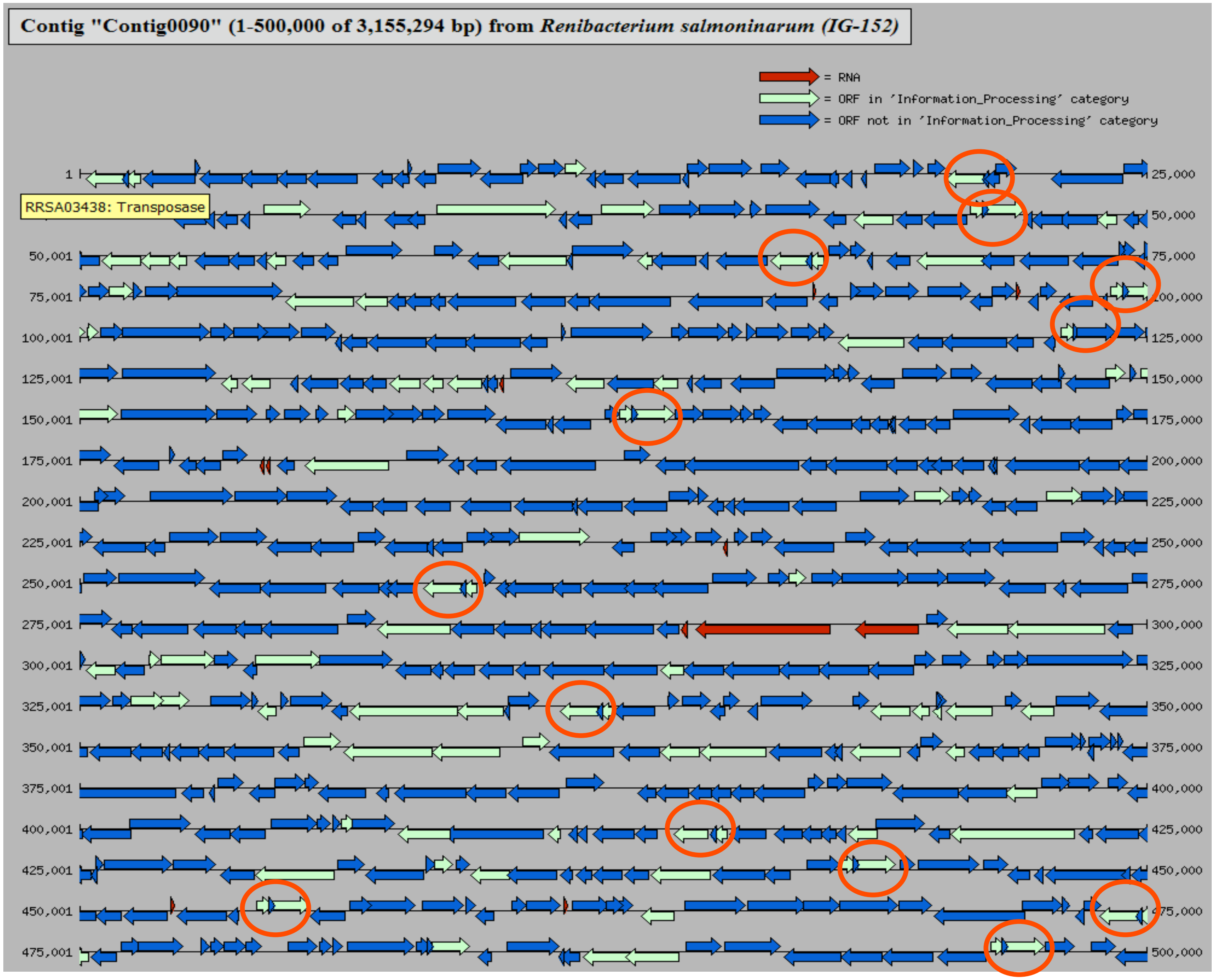
Renibacterium: genome map



Mapped Transposase ORFs

© 2005 Integrated Genomics, Inc.

Renibacterium: 'transposases'



Contig "Contig0090" (1-500,000 of 3,155,294 bp) from *Renibacterium salmoninarum (IG-152)*

🟥 = RNA

🟩 = ORF in 'Information_Processing' category

🟦 = ORF not in 'Information_Processing' category

RRSA03438: Transposase

© 2005 Integrated Genomics, Inc.

Renibacterium: central metabolism

R. salmoninarum possesses classical metabolic pathways for:

- glycolytic pathway (EMP)
- pentose phosphate
- TCA (Krebs) cycle
- pyruvate cycle

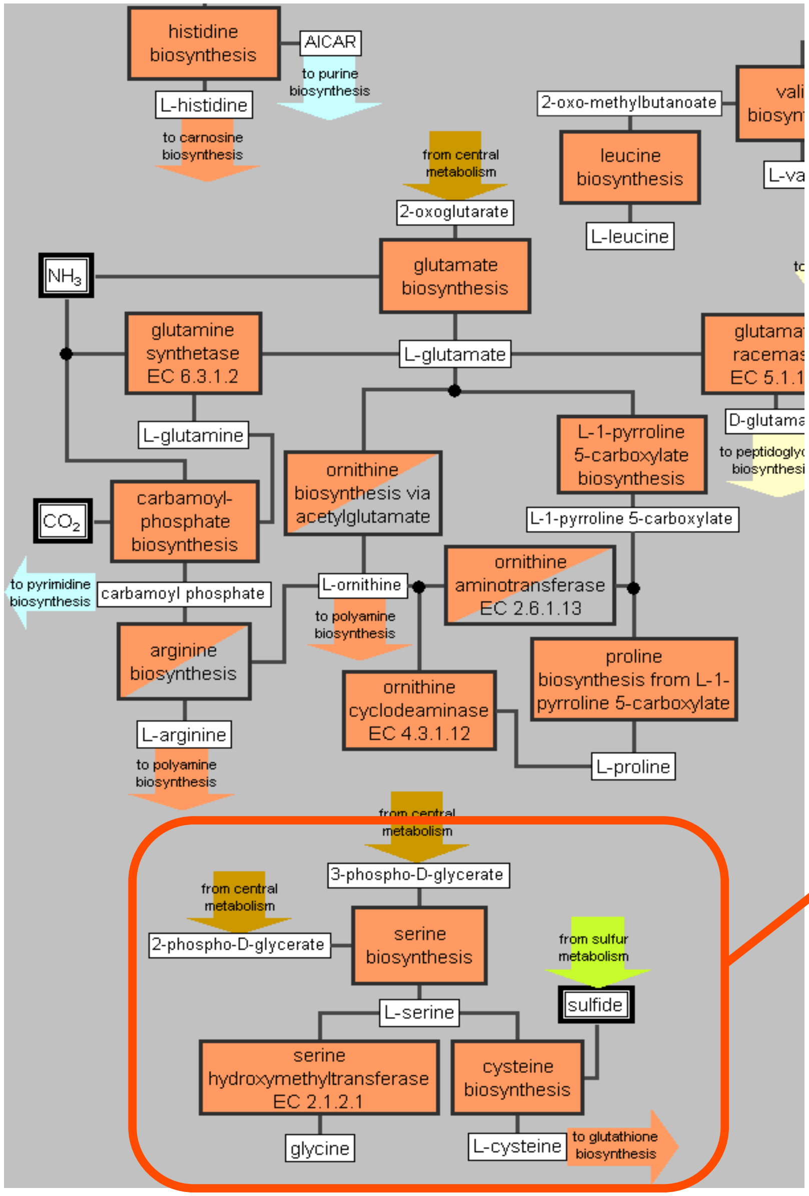
Summary: can utilize many sugars and polyols

R. salmoninarum has limited transporters for sugars/polyols:

- glucose/fructose
- fructose PTS
- glycerol
- gluconate
- arabinose
- C4 dicarboxylate (malate/succinate)

Summary: most likely able to uptake at least fructose, gluconate, glycerol

Renibacterium: amino acid biosynthesis



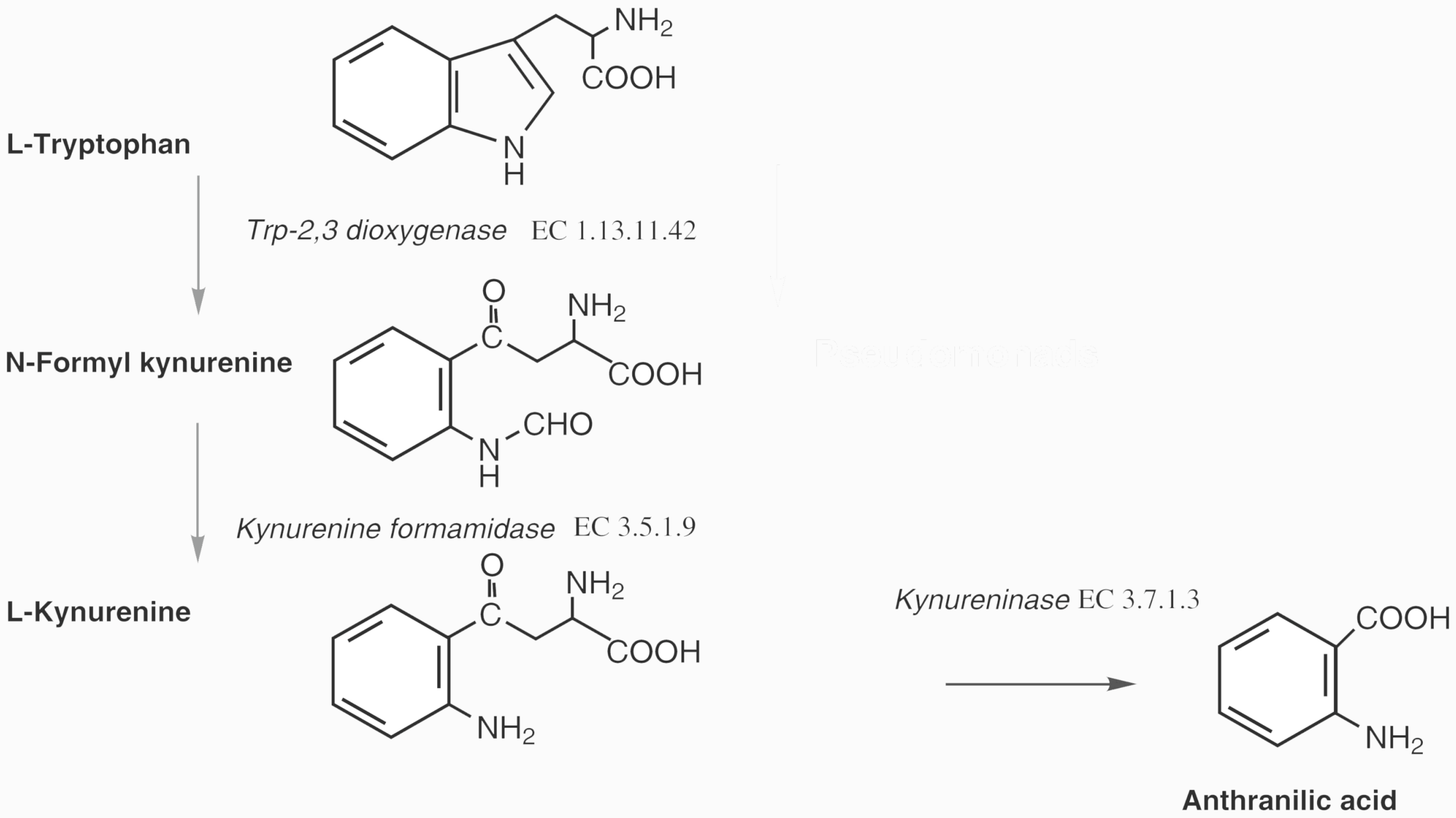
Initial inspection suggests that the bacterium is able to make most amino acids

Question: Bacterium can biosynthesize serine and cysteine. So why is cysteine added to KDM2 medium for *Rsa* growth?

KDM2 media:

yeast extract, peptone, serum or charcoal, 0.05% Cys

Renibacterium: what is L-kynurenine?



L-Tryptophan

Trp-2,3 dioxygenase EC 1.13.11.42

Pseudomonads

N-Formyl kynurenine

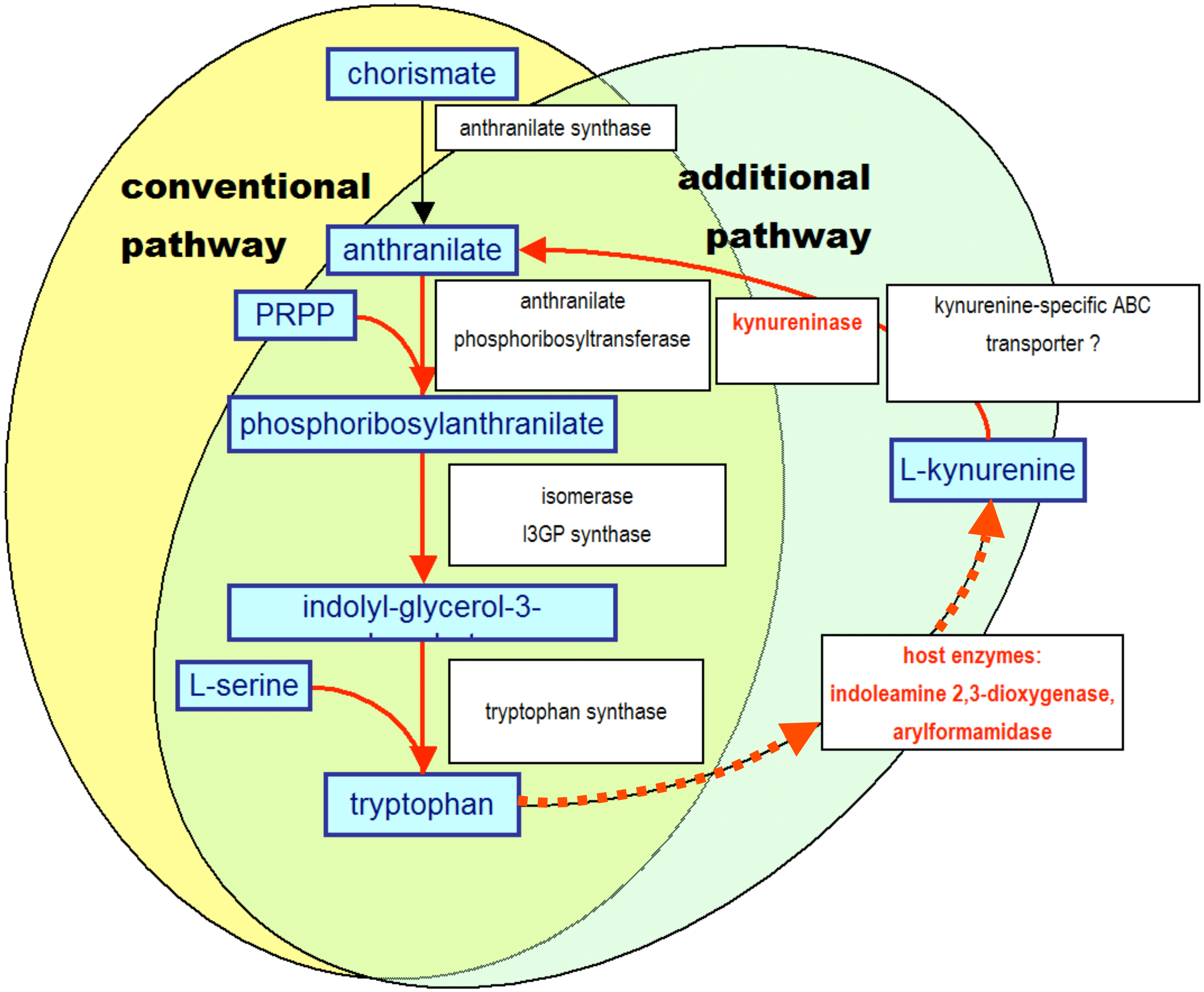
Kynurenine formamidase EC 3.5.1.9

L-Kynurenine

Kynureninase EC 3.7.1.3

Anthranilic acid

Renibacterium: scavenging host kynurenine?



Renibacterium: iron uptake and oxygen stress

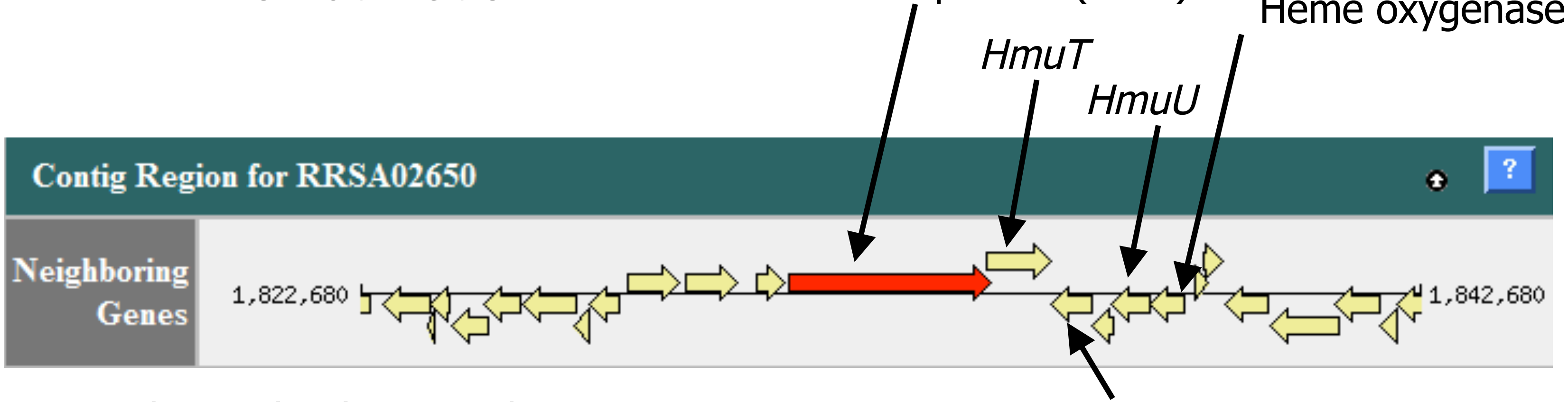
- Iron utilization

Heme transport associated protein (*HtaA*)

HmuT

HmuU

Heme oxygenase



HmuV

HmuT: hemin-binding periplasmic protein
HmuU: hemin transport system permease protein
HmuV: hemin transport system, ATP-binding protein

- Enzymes capable of coping with intracellular oxygen stress
  - superoxide dismutase (RRSA02650)
  - catalase (RRSA00708)
  - thioredoxin peroxidase (RRSA01668)

Renibacterium: pathogenicity factors?

Major cell surface antigen (Msa1, p57)

- Msa1 (RRSA00268, RRSA03467)
- attachment to salmonid erythrocytes

Cell adhesion protein

- hemolysin *tlyA*-related, RRSA01172

Metalloprotease/Hemolysin

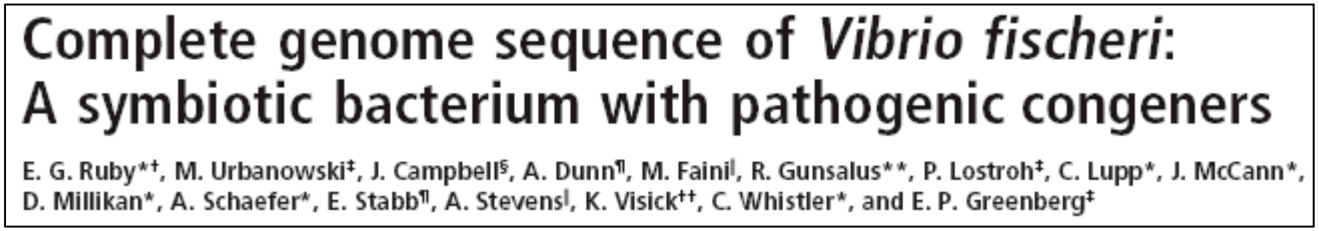
- known *Rsa hly* (maps to RRSA01873)

Virulence factor MviN- and MviB-like (*S. typhimurium*)

Chitin and chitosan hydrolyzing enzymes identified

- endochitinase (RRSA06681)
- chitosanase (RRSA00952)
- β-N-acetylhexosaminidase (RRSA01602)
- capable of utilizing N-acetylglucosamine, D-glucosamine and chitobiose

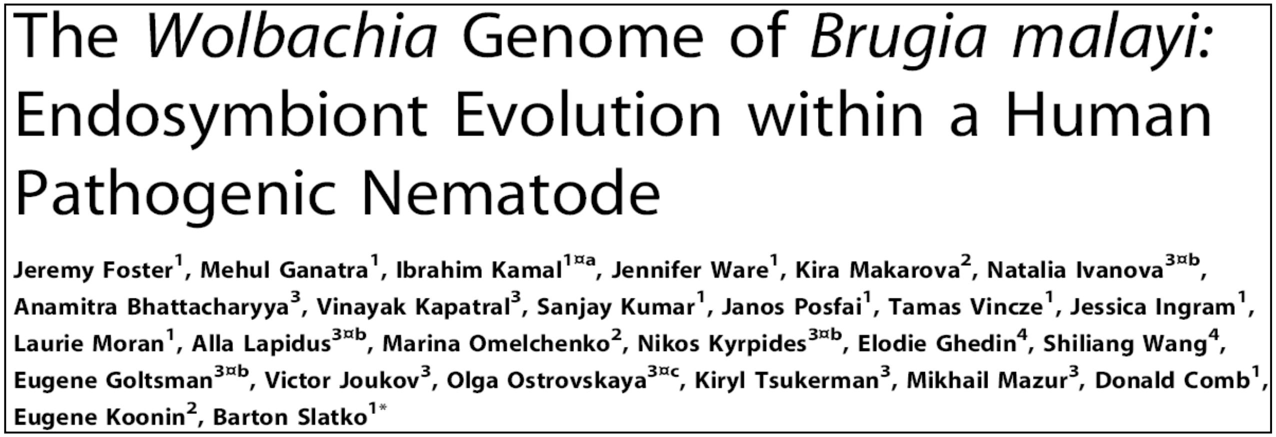
ERGO: selected publications



Complete genome sequence of *Vibrio fischeri*: A symbiotic bacterium with pathogenic congeners

E. G. Ruby[*†], M. Urbanowski[‡], J. Campbell[§], A. Dunn[¶], M. Faini[‖], R. Gunsalus[**], P. Lostroh[‡], C. Lupp[*], J. McCann[*], D. Millikan[*], A. Schaefer[*], E. Stabb[¶], A. Stevens[‖], K. Visick[††], C. Whistler[*], and E. P. Greenberg[‡]

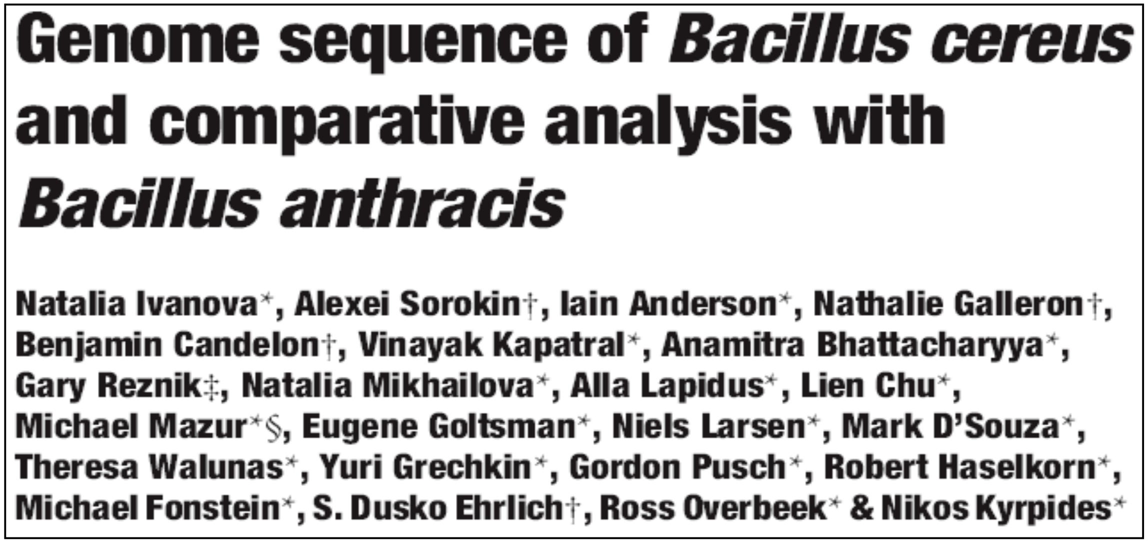
PNAS (2005)



The *Wolbachia* Genome of *Brugia malayi:* Endosymbiont Evolution within a Human Pathogenic Nematode

Jeremy Foster[1], Mehul Ganatra[1], Ibrahim Kamal[1¤a], Jennifer Ware[1], Kira Makarova[2], Natalia Ivanova[3¤b], Anamitra Bhattacharyya[3], Vinayak Kapatral[3], Sanjay Kumar[1], Janos Posfai[1], Tamas Vincze[1], Jessica Ingram[1], Laurie Moran[1], Alla Lapidus[3¤b], Marina Omelchenko[2], Nikos Kyrpides[3¤b], Elodie Ghedin[4], Shiliang Wang[4], Eugene Goltsman[3¤b], Victor Joukov[3], Olga Ostrovskaya[3¤c], Kiryl Tsukerman[3], Mikhail Mazur[3], Donald Comb[1], Eugene Koonin[2], Barton Slatko[1*]

PLoS Biology (2005)



Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*

Natalia Ivanova[*], Alexei Sorokin[†], Iain Anderson[*], Nathalie Galleron[†], Benjamin Candelon[†], Vinayak Kapatral[*], Anamitra Bhattacharyya[*], Gary Reznik[‡], Natalia Mikhailova[*], Alla Lapidus[*], Lien Chu[*], Michael Mazur[*§], Eugene Goltsman[*], Niels Larsen[*], Mark D'Souza[*], Theresa Walunas[*], Yuri Grechkin[*], Gordon Pusch[*], Robert Haselkorn[*], Michael Fonstein[*], S. Dusko Ehrlich[†], Ross Overbeek[*] & Nikos Kyrpides[*]

Nature (2003)

Future directions

- Metagenomics: analysis of mixed microbial communities
- Affordable, rapid sequencing and analysis (typing) of entire strain collections
- Integration of other data types *e.g.* phenotype microarrays
- Molecular diagnostics: expression chips and probe design

Acknowledgements

***Xylella / Fusobacterium* comparative genomics**

Bill and Helene Feil (UC Berkeley)

Joint Genome Institute (DoE)

Vinayak Kapatral (Integrated Genomics)

Renibacterium salmoninarum

Mark Strom (NOAA)

Greg Wiens (USDA)

Henry Burd (Integrated Genomics)

Select IG Customers

Corporate

ADM
BASF
Cargill
Christian-Hansen
Danone
Degussa
Diversa
Dow
Dow AgroSciences
Kyowa Hakko
Genencor
Nestle
New England Biolabs
Pfizer
Proctor & Gamble
Roche

Strategic Partners

Agencourt Biosciences (U.S.)
GATC (Germany)

Governmental

RML/NIAID/NIH
US Air Force
US Army
US Department of Defense
USDA
NOAA

Academic (various)

Univ of Chicago
Univ of California
Ohio State
Utah State
Univ of Florida
Univ of Wisconsin
Univ Wageningen (NL)
Univ of Goettingen